# Benford's Law and Power Law Behavior in Fragmentation Processes

BY

Joseph R. Iafrate

WITH

Frederick W. Strauch and Steven J. Miller, Advisors


A thesis submitted in partial fulfillment
of the requirements for the
Degree of Bachelor of Arts with Honors
in Physics

WILLIAMS COLLEGE
Williamstown, Massachusetts

May 26, 2014

# Abstract

We construct and analyze models for the fragmentation of a conserved quantity. Using a statistical model, we derive an approximation as well as bounds for the restricted partitioning problem, which we apply to the distribution of fragments. We also modify the canonical ensemble from statistical physics. Taken together, we set a threshold on the magnitude of the conserved quantity needed to result in power law behavior, as well as a threshold on the number of possible piece sizes in a special case. We also investigate variations on two specific fragmentation procedures, the directed and undirected fragmentations, for power law behavior. Calculations show that the undirected fragmentation exhibits power law behavior. We consider small perturbations to process rates and find that the multi-path fragmentation is affected less as the number of piece sizes grows. We confirm this numerical result using first-order perturbation theory.

# Acknowledgements

In order to thank everyone who assisted in the grand adventure that was this thesis, I would need to list anyone who has ever encouraged my love of physics or supported me in my academic endeavors. This support has ranged from encouraging me to sleep to asking for an actual explanation of my thesis project (to you, I say "God bless you"). That said, there are a few individuals I would be remiss not to mention by name.

First and foremost, I want to thank my mom for her support from day one, not just of this thesis but of my academic pursuits. She has always encouraged me to try my hardest and to follow my dreams, no matter where they led. It is because of her that I made it this far.

Next, I would like to thank my fellow physics major and basement dweller Nathan Bricault. Whenever *Mathematica* refused to comply, Nathan was always there to talk me through the struggle. It was comforting to know there was always someone in the lab down the hall that I could visit when I needed a thesis break. Most importantly, I would never have succeeded as a physics major without our collaborations.

None of this work would have been possible without the guidance of my two advisors, Professors Steven J. Miller and Frederick W. Strauch. I have been incredibly lucky to have taken multiple courses with both of these men over my junior and senior years. This has been such a unique experience, being able to work on a project straddling the line between mathematics and physics. I am convinced that most of my mathematical intuition has come from Professor Miller (I immediately have logs ready to go whenever I see a product). This is fitting, due to the mathematical opportunities he has given me, through classes, collaboration with SMALL 2013, and this project. One day I hope we can agree on an MVP winner. Professor Strauch has been a calming presence in what became an increasingly hectic senior year. He has been caring and understanding while still giving me the freedom to explore the work on my own. I am deeply grateful for his mentorship, both in research and in determining my post-graduate plans. I hope his enthusiasm for his research has rubbed off on me.

Finally, I must thank the One who has given me the opportunity to attend Williams and to study the world He has created. He has blessed me richly. Christ has been my rock throughout my time at Williams and will be wherever my physics career takes me next.

# Contents

# List of Figures

# Chapter 1

# Introduction

*"Why are there more small things in the world than large?"*

So begins Don Lemons' 1985 paper *On the numbers of things and the distribution of first digits* [11]. Though this seems a philosophical musing more than anything else, it was not without motivation. As we will see, a varied collection of real-world data sets seem to follow the same peculiar pattern. In each, one would find a majority of the data entries to start with a 1, with decreasing probabilities as the first digit grew. Many mathematicians and physicists have attempted to divine the source of this strange behavior, but some simply dismiss it as a quirk of the number system. Lemons saw otherwise. He could not answer "Why?", nor did he claim to. Instead, his answer to "When?" has the potential to impact many fields of science.

## 1.1 Benford's Law

Lemons was not the first to identify this strange pattern. In 1881, mathematician Simon Newcomb noted that the first few pages of logarithm tables tended to be more worn than the pages that followed [14]. That is, a scientist referencing the tables would more likely need the logarithm of a number that started with a 1 than with a 2, and so on. Newcomb interpreted this to mean that numbers starting with a first digit of 1 occurred more often than numbers with a different starting digit.

In 1938, physicist Frank Benford made the same observation [5]. Benford compiled a list of over 20,000 observations that seemed to follow the same pattern. Among these were population counts, the area of rivers, specific heats, addresses, and death rates. He concluded, like Newcomb had years before, that the frequency

of first digits followed the logarithmic relation

$$\Pr\left(\text{first digit } m\right) = \log_{10} \frac{m+1}{m}. \tag{1.1}$$

Today this is known as Benford's Law. Benford also derived expressions for the probability of an arbitrary digit, but the first digit has received the majority of attention. The Benford probabilities are plotted in Figure 1.1.

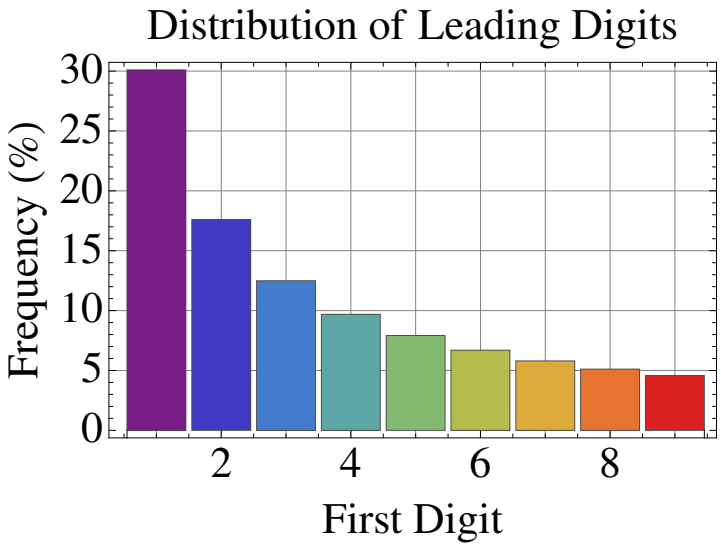



Figure 1.1: The probabilities of each possible first digit for a data set that follows Benford's Law.

We say that a data set is Benford if the distribution of its first digits follow Benford's Law. Other examples of Benford data sets include the Fibonacci numbers, the total number of print materials in US libraries, and Twitter users by their follower count [12].

Benford's Law is not a quirk of some particular counting system. It is an invariant distribution. That is, if a data set is Benford, it is still Benford even if the units or scale is changed. For example, if we measured area in acres rather than square miles, the river areas from Benford's paper would still have the same first-digit distribution. Pinkham demonstrated that if a distribution of first digits existed that was scale invariant, it had to be Benford's Law [16]. Furthermore, Benford behavior is not unique to data sets in base 10. Benford's Law generalizes to any base $b$ (though a data set that is Benford in base 10 is not necessarily Benford in another base) [9]. Many have proposed explanations for the prevalence of Benford in real-world data sets. However, there is no consensus among mathematicians regarding these [6]. Perhaps the most promising explanation was

proposed by Hill [10]. He showed that random samples drawn from a set of random distributions followed Benford's Law. For a more in-depth introduction to Benford's Law, see [20].

## 1.2 On the Number of Things

The goal of Lemons' paper was to determine in which situations a data set may be Benford. He presented a model for fragmenting a conserved quantity. Given some quantity $X$, consider every possible way to break it up into pieces between specified upper and lower bounds, assuming each distribution of pieces equally probable. Then, Lemons claimed, on average the number of pieces of a given size is inversely proportional to the piece size. In the limit of continuous piece sizes, this power law results in the pieces being Benford distributed. We treat Lemons' argument in detail in Section 2.1. Because Benford's Law emerges from this model, Lemons makes the bold claim that whenever a conserved quantity is fragmented, we expect its piece sizes to follow Benford's Law. This assertion applies to a large number of quantities studied by physicists, among them mass, energy, and volume, so the veracity of the Lemons result has a bearing on many data sets we work with regularly.

There are, however, problems with Lemons' paper. First, the way he defines the piece sizes in his fragmentation is unclear. He allows the piece sizes to be any real number, giving a continuous spectrum of piece sizes and forcing a binning which he does not seem to standardize. Second, he does not specify any constraints on the size of the conserved quantity or the number of possible piece sizes. We suspect that these play a role in the emergence of power law behavior, so that Lemons' result is not as universal as he puts forth.

## 1.3 Our Work

It was this skepticism of Lemons' paper that led to the inception of this project. In an attempt to confirm his result, we set out to redefine Lemons' model from scratch and show under what conditions a power law distribution (and hence Benford behavior) is a consequence.

In real-world fragmentation processes, there is often some base unit of which all pieces are multiples. For instance, in a nuclear fragmentation, the nucleus breaks into integer numbers of nucleons. We do not see fractional nucleons. Hence we model fragmentation of an integral conserved quantity into integer pieces. We derive a formula for the average number $\langle n_j \rangle$ of pieces of size $x_j$ from our model using a variety of mathematical methods. Then we determine the range of $X$ for

which Lemons' result holds. For our statistical mode, we find that the magnitude of the conserved quantity $X$ must be much larger than $N^2 x_N$, where $N$ is the number of possible piece sizes and $x_N$ is the largest piece size. We also examine a model based on the canonical ensemble distribution of statistical physics. This model shows power law behavior for $X \gg N x_N$. The derivation of these results as well as a comparison of our models and methods comprise a majority of Chapter 2.

To get from power law to Benford, Lemons took the piece sizes to be continuous. We may not do so. Instead, we calculate in a particular fragmentation scheme how large our set of possible piece sizes must be in order for the first-digit probabilities to be close to Benford. This is addressed in Section 2.3. For a set of consecutive integer piece sizes, we have that the first digit probabilities are within 20% of Benford by $N \sim 10^4$ and within 10% by $N \sim 10^9$. Both quoted error percentages are relative error with respect to Benford.

We believe that in some cases the fragmentation process can affect the likelihood of resulting piece distributions. Thus we analyze variants of two specific processes, single-path and multi-path fragmentations, and determine under what specifications those also lead to a power law distribution. Having identified that a multi-path fragmentation will result in a power law when the break-up rate exactly counters the recombination rate in Section 3.1, we see how far we can push that result in Section 3.2. That is, we calculate how far we may perturb our processes from the ideal while still preserving the desired outcome, giving us a sense of the robustness of the power law distribution. We find that the multi-path undirected fragmentation stays power law for a range of perturbations and becomes more impervious to fluctuations as more piece sizes are considered.

All of these investigations are undertaken with a mind towards identifying Benford behavior in the world around us.

# Chapter 2

# Approaching Benford's Law - Convergence to the Distribution

## 2.1 The Lemons Scenario

Lemons [11] argued that the pieces resulting from a fragmentation process of a conserved quantity should, on average, follow Benford's Law. We repeat his argument in brief here.

Consider a conserved quantity of size $X$. By some process, it is broken into $n_j \Delta x_j$ pieces with magnitudes between $x_j$ and $x_j + \Delta x_j$, where $j = \{1, 2, \ldots, N\}$. He defines $N$ as the the number of divisions between two fixed bounds $x_l$ and $x_u$ (lower and upper, respectively). Put another way, Lemons sorts the resulting pieces into $N$ bins, where the $j$th bin is of size $\Delta x_j$. The base value of each bin must fall within a prescribed range. Using the conservation relation

$$X = \sum_{i=1}^{N} x_j n_j \Delta x_j \tag{2.1}$$

and simple arguments regarding the averaging of random variables, he finds that the average number of pieces of a given size $x_j$ is

$$\langle n_j \rangle = \frac{X}{N x_j \Delta x_j}, \tag{2.2}$$

where $n_j$ and $x_j$ are discrete. In the limit that the bin sizes are equivalent and $\Delta x_j \to 0$, Lemons concludes that $\langle n_j \rangle \propto \frac{1}{x}$. This limit is analogous to assuming continuous piece sizes. A distribution of this form leads directly to Benford's Law for the probability of the first digit:

$$\Pr(\text{first digit } m) \;=\; \int_m^{m+1} \frac{dx}{x} \Big/ \int_{10^p}^{10^{p+1}} \frac{dx}{x}$$

$$
\begin{aligned}
&= \log \frac{m+1}{m} / \log 10 \\
&= \log_{10} \frac{m+1}{m}. \qquad (2.3)
\end{aligned}
$$

Because Lemons did not assume any particular fragmentation scheme, he concludes that the pieces of any conserved quantity will follow Benford's Law in the average. In physics, this conclusion has far-reaching effects, as many important quantities (among them mass, energy, and momentum) are conserved in familiar scenarios.

While Lemons' result is impressive in its simplicity, it is not entirely clear how he defines his piece size bins. We may assume that the $x_j$ are spaced such that $x_j + \Delta x_j = x_{j+1}$, so that the domain is entirely split into the $j$ bins, but that is just our interpretation. Furthermore, if $\Delta x_j$ is the bin size, we do not see why it also determines the total number of pieces produced.

These confusions aside, we observe that this model does not apply to relevant physical systems. For instance, consider the case of nuclear fragmentation, as investigated by Chase and Mekjian [7]. For these problems, it is assumed that the nuclei are composed of an integer number of nucleons (further internal structure is usually ignored). Even if it were possible to obtain a fractional part of a nucleon, there is a limit to how small a fraction we can measure. Real-world measurements are effectively discrete and introduce some minimum gradation (having a finite number of digits). While the continuous piece size limit simplifies calculations, it does not necessarily reflect reality. However, as in most physics problems, we would like to determine when it is a valid approximation.

We propose a model based on Lemons' model that only allows for integer piece sizes given an integral conserved quantity. This model has two variable parameters, the magnitude of the conserved quantity $X$ and the number of piece sizes $N$, and the integer piece sizes $x_j$ are allowed to vary. We want to determine in what regimes or under which assumptions our model reaches the same limiting behavior as Lemons' - that is, for what relations of $X$ and $N$ do we expect our average piece sizes to follow a power law distribution? Specifically, we want to know when

$$\langle n_j \rangle = \frac{X}{N x_j} \qquad (2.4)$$

holds. Once known, we will consider how Benford's Law materializes from this model. This task is the subject of the following sections.

In this chapter, we establish the terminology and assumptions we will use in our model. We use multiple techniques to place restrictions on $X$. Then, we draw upon statistical mechanics to derive a different model. After comparing these results, we use a numerical simulation to get a sense of how large $N$ must be for power law behavior to result in Benford probabilities.

## 2.2 Variable Conserved Quantity - The Power Law

### 2.2.1 A Statistical Model

We cast our scenario as a restricted partition problem over the integers. That is, we consider an integer decomposing into a collection of other integers such that the sum of the members of that collection is equivalent to the starting integer.

Given an integer $X$ and set of integers $H = \{x_1, x_2, \ldots, x_N\}$, we can partition $X$ such that

$$X = \sum_{\substack{j=1 \\ x_j \in H}}^{N} n_j x_j. \tag{2.5}$$

This is called a *restricted partition of* $X$ [2]. We denote the number of restricted partitions of $X$ into parts from $H$ by $P_H(X)$. For reference, $X$ corresponds to our conserved quantity, and $H$ is the set of all potential piece sizes, ordered such that $x_1 < x_2 < \cdots < x_N$. Eq. (2.5) is the conservation relation for our fragmentation process, and $P_H(X)$ gives all possible fragmentation outcomes. Our desired quantity is $\langle n_j \rangle$, the average number of pieces of size $x_j$. To calculate it, we need an expression for the total number of parts $x_j$ in all the restricted partitions of $X$ as well as an expression for the total number of partitions $P_H(X)$. The quotient of these two expressions yields the desired average.

The generating function for the number of restricted partitions [2] is

$$\sum_{k=0}^{\infty} P_H(k) q^k = \prod_{x \in H} (1 - q^x)^{-1}, \tag{2.6}$$

with $q < 1$ to ensure convergence.

We can think of each term in the product $(1 - q^{x_j})^{-1}$ as the closed-form expression for the geometric series $\sum_{n=0}^{\infty} (q^{x_j})^n$. Then the right-hand side of Eq. (2.6) is a product of geometric series, resulting in an infinite sum of various powers of $q$. We can group common powers of $q$ together so that the right-hand side becomes a power series in $q$. Then the coefficient of $q^X$ is $P_H(X)$ by comparison with the left-hand side. To justify this interpretation, we observe that $q^X$ only occurs on the right-hand side if $X$ can be written as a linear combination of the $x_j$ (as we are multiplying together $q$ with powers that are multiples of the $x_j$). We get $q^X$ for every possible combination. That number of combinations is just $P_H(X)$. To isolate the desired quantity, we differentiate Eq. (2.6) with respect to $q$ a number of times equal to $X$ (acquiring an unwanted factor of $X!$ along the way). Since $P_H(X)$ is the constant term in a power series in $q$, we can isolate it by setting $q = 0$.

However, we want to count the total number of times a particular $x_j$ occurs in our partitions. We return to Eq. (2.6) and differentiate with respect to $x_j$. Thinking again of the right-hand side as a product of geometric series, we see that this multiplies each term by the number of $x_j$ in the partition it represents. This also results in an extra factor of $\log q$. When we group $q$ by common powers, we can differentiate $X$ times with respect to $q$ to isolate $P_H(X)$ as before. But now our quantity is multiplied by the number of times $x_j$ appeared in each exponential argument that totaled $X$. This is the total number of $x_j$ that occurred in all partitions of $X$.

We set $q = 0$ to discard the other terms. Thus the total number $n_j$ in all partitions is given by

$$\frac{1}{X!} \times \left(\frac{\partial}{\partial q}\right)^X \left[\frac{1}{\log q} \frac{\partial}{\partial x_j} \prod_{x \in H} (1 - q^x)^{-1}\right]_{q=0}. \tag{2.7}$$

We can simplify this expression, as

$$\begin{aligned} \frac{\partial}{\partial x_j} \prod_{x \in H} (1 - q^x)^{-1} &= \left(\frac{\partial}{\partial x_j} \frac{1}{1 - q^{x_j}}\right) \prod_{x \in H \setminus \{x_j\}} (1 - q^x)^{-1} \\ &= \frac{q^{x_j}}{(1 - q^{x_j})^2} \log q \prod_{x \in H \setminus \{x_j\}} (1 - q^x)^{-1} \\ &= \frac{q^{x_j}}{1 - q^{x_j}} \log q \prod_{x \in H} (1 - q^x)^{-1}. \end{aligned} \tag{2.8}$$

Furthermore, we use Leibniz' theorem for differentiation of a product [1]:

$$\begin{aligned} &\left(\frac{\partial}{\partial q}\right)^X \left[\frac{q^{x_j}}{1 - q^{x_j}} \left(\prod_{x \in H} (1 - q^x)^{-1}\right)\right]_{q=0} \\ &= \sum_{k=0}^{X} \binom{X}{k} \left[\frac{\partial^k}{\partial q^k} \left(\frac{q^{x_j}}{1 - q^{x_j}}\right)\right] \left[\frac{\partial^{X-k}}{\partial q^{X-k}} \left(\prod_{x \in H} (1 - q^x)^{-1}\right)\right] \\ &= \sum_{k=0}^{X} \binom{X}{k} \left[\frac{\partial^k}{\partial q^k} \left(\frac{q^{x_j}}{1 - q^{x_j}}\right)\right] (X - k)! P_H(X - k) \\ &= \sum_{k=0}^{X} \frac{X!}{k!} \left[\frac{\partial^k}{\partial q^k} \left(\frac{q^{x_j}}{1 - q^{x_j}}\right)\right] P_H(X - k), \end{aligned}$$

where we recognize the $(X - k)$th derivative of the generating function Eq. (2.6) as $(X - k)! P_H(X - k)$. Note that $q$ is implicitly being set to 0 on the right side of

the equation and will be in what follows, so we must make use of one last equality before we can set $q = 0$ across the equals sign. Thus, Eq. (2.7) is equivalent to

$$\sum_{k=0}^{X} \frac{1}{k!} \left[ \frac{\partial^k}{\partial q^k} \left( \frac{q^{x_j}}{1 - q^{x_j}} \right) \right] P_H(X - k). \tag{2.9}$$

Now, one can show using the geometric series formula that

$$\frac{\partial^k}{\partial q^k} \left( \frac{q^{x_j}}{1 - q^{x_j}} \right) = \frac{\partial^k}{\partial q^k} \left( \sum_{l=0}^{\infty} q^{x_j(l+1)} \right). \tag{2.10}$$

If $k$ is a multiple of $x_j$, setting $q = 0$ evaluates the above expression to $k!$. If it is not, the derivatives will have no constant terms and vanish when we set $q = 0$. Thus we have

$$\sum_{k=0}^{X} \frac{1}{k!} \left[ \frac{\partial^k}{\partial q^k} \left( \frac{q^{x_j}}{1 - q^{x_j}} \right) \right]_{q=0} P_H(X - k) = \sum_{i=0}^{\lfloor X/x_j \rfloor} P_H(X - ix_j), \tag{2.11}$$

where we have made the replacement $\frac{k}{x_j} \to i$. We divide by $P_H(X)$ to obtain the desired result:

$$\langle n_j \rangle = \frac{1}{P_H(X)} \sum_{i=0}^{\lfloor X/x_j \rfloor} P_H(X - ix_j). \tag{2.12}$$

Provided we can compute $P_H(X)$, this compact expression allows us to find the desired quantity.

It turns out that we may represent $P_H(X)$ as a sum over our piece sizes. To justify this, consider the following analogy. The physicist from the lab next door asks that we give him change for \$1 in pennies, nickels, dimes, and quarters. Counting all the possible combinations can be accomplished by summing over the allowed numbers of each coin denomination, which we represent as $\{n_{.01}, n_{.05}, n_{.10}, n_{.25}\}$. First we determine how many quarters go into \$1, the answer being 4. That's one option: $\{0, 0, 0, 4\}$. Now consider 1 fewer quarter, so we only use 3. We still have to account for \$0.25, so we move on to dimes and ask how many dimes we may give. We may only use 2, as 3 would put us over \$1. This leaves \$0.05, which we can satisfy with a nickel. This gives us another combination: $\{0, 1, 2, 3\}$. Now we go back one step and decrease the number of nickels, making the difference up with pennies: $\{5, 0, 2, 3\}$. We can continue this until we have considered every possible combination of coins we could give in change.

To formalize this procedure, we return to our general model. We start by considering how many pieces of our largest size, $x_N$, may result from a partition of $X$. Selecting one such number, we then consider how many pieces of size $x_{N-1}$

can result from a partition of the "remainder" integer, $X - n_N x_N$. We select one such number $n_{N-1}$ and continue. Once we have determined the number of pieces of each size to include, we check that our conservation relation Eq. (2.5) still holds. If so, we count the partition. If not, we discard it. What we are actually summing is a Kronecker delta function whose argument is the difference between the total of our pieces and $X$. Now we can formally write our expression as

$$P_H(X) = \sum_{n_N=0}^{\lfloor L_N \rfloor} \sum_{n_{N-1}=0}^{\lfloor L_{N-1} \rfloor} \cdots \sum_{n_2=0}^{\lfloor L_2 \rfloor} \sum_{n_1=0}^{\lfloor L_1 \rfloor} \delta \left( X - \sum_{h \in H} n_h x_h \right), \tag{2.13}$$

where the upper limits $L_k$ are given by

$$L_k = \frac{X - \sum_{l=k}^{N} n_l x_l}{x_k}, \tag{2.14}$$

and $\lfloor L_k \rfloor$ is the floor or $L_k$ (the largest integer less than or equal to $L_k$).

For $H$ of small cardinality and small $X$, we can compute this quantity directly or with the aid of computational software. However, a closed-form expression for $P_H(X)$ without sums does not exist, though progress continues to be made in the mathematical literature [3]. We have the techniques to either approximate it or bound it in terms of other expressions. We pursue both approaches in the following sections.

### 2.2.2 Obtaining $\langle n_j \rangle$ by Approximating $P_H(X)$

Our first approach is to approximate $P_H(X)$ to sufficient order in $X$. We may then substitute that expression into our formula for $\langle n_j \rangle$. In the following calculations, we must make use of the following lemma. The proof is left to the reader.

**Lemma 2.2.1** Power Sum
*A finite sum of integers, each raised to the power p, can be bounded in the following way:*

$$\frac{1}{p+1} n^{p+1} < \sum_{k=1}^{n} k^p < \frac{1}{p+1} n^{p+1} + n^p. \tag{2.15}$$

*Furthermore, we may approximate the sum by taking the average of the bounds:*

$$\sum_{k=1}^{n} k^p \approx \frac{1}{p+1} n^{p+1} + \frac{1}{2} n^p. \tag{2.16}$$

We start with the full expression for $P_H(X)$, Eq. (2.13). We restrict our attention to those sets $H$ such that $x_1 = 1$. This ensures that a partition exists for every $X$ and allows us to sum over the delta function. Furthermore, this is in line with physical systems that are usually defined in terms of some base element (like a nucleon). This lets us disregard the sum over $n_1$, as once the other $n_j$ are determined, it only has one possible value. Now, the floor functions on the $L_j$ in the upper sum limits ensure that those limits are all integers. Our first approximation is to remove those floors. Note that $\lfloor L_j \rfloor \leq L_j$ for all $j$, so we are overcounting. With the above replacements, Eq. (2.13) becomes

$$P_H(X) \approx \sum_{n_N=0}^{L_N} \sum_{n_{N-1}=0}^{L_{N-1}} \cdots \sum_{n_2=0}^{L_2} 1. \tag{2.17}$$

Notice that we can rewrite $L_{N-1}$ as $\frac{x_N}{x_{N-1}}(L_N - n_N)$. We thus define a recurrence relation for all the $L_j$:

$$L_j = \frac{x_{j+1}}{x_j}(L_{j+1} - n_{j+1}), \qquad j = \{2, 3, \ldots, N-1\}. \tag{2.18}$$

Now we can evaluate each individual sum, starting with the sum over $n_2$ and proceeding to larger $j$. Our second approximation will be to to drop all terms except the two highest orders in $X$ after the evaluation of each summation. Often, $X$ will be larger than the piece sizes involved (though we will quantify this later), so we may assume that $X^j \gg X^{j-2}$ and that we may drop the latter and any smaller terms. By their very definition, each $L_j$ is roughly proportional to $X$, so this approximation is equivalent to keeping only the largest two orders in $L_j$.

We now evaluate. In the calculations that follow, we refer to the sum over $n_j$ as the $j$th sum for convenience even though it will be the $(j-1)$th sum evaluated. Using Eqs. (2.16) and (2.18), we see that

$$\begin{aligned}
\sum_{n_N=0}^{L_N} \cdots \sum_{n_3=0}^{L_3} \sum_{n_2=0}^{L_2} 1 &= \sum_{n_N=0}^{L_N} \cdots \sum_{n_3=0}^{L_3} (1 + L_2) \\
&= \sum_{n_N=0}^{L_N} \cdots \sum_{n_3=0}^{L_3} \left[1 + \frac{x_3}{x_2}(L_3 - n_3)\right] \\
&= \sum_{n_N=0}^{L_N} \cdots \left[\sum_{n_3=0}^{L_3} (1 + \frac{x_3}{x_2} L_3) - \frac{x_3}{x_2} \sum_{n_3=0}^{L_3} n_3\right] \\
&= \sum_{n_N=0}^{L_N} \cdots \sum_{n_4=0}^{L_4} \left[(1 + \frac{x_3}{x_2} L_3)(1 + L_3) - \frac{x_3}{x_2}\left(\frac{1}{2}L_3^2 + \frac{1}{2}L_3\right)\right]
\end{aligned}$$

$$= \sum_{n_N=0}^{L_N} \cdots \sum_{n_4=0}^{L_4} \left[ 1 + L_3 + \frac{x_3}{x_2} L_3 + \frac{x_3}{x_2} L_3^2 - \frac{x_3}{2x_2} L_3^2 - \frac{x_3}{2x_2} L_3 \right]$$
$$= \sum_{n_N=0}^{L_N} \cdots \sum_{n_4=0}^{L_4} \left[ \frac{x_3}{2x_2} L_3^2 + \frac{2x_2 + x_3}{2x_2} L_3 + 1 \right].$$

We drop the lowest order term after evaluating the 3rd sum:

$$\sum_{n_N=0}^{L_N} \cdots \sum_{n_4=0}^{L_4} \left[ \frac{x_3}{2x_2} L_3^2 + \frac{2x_2 + x_3}{2x_2} L_3 \right]. \tag{2.19}$$

Rather than continue through all $N-1$ sums, we use induction to reach the final expression. We claim that after the $j$th sum, we drop all terms ($L_j^{j-3}$) and lower to obtain an expression of the form

$$\sum_{n_N=0}^{L_N} \cdots \sum_{n_{j+1}=0}^{L_{j+1}} \left[ \frac{x_j^{j-2}}{(j-1)! D_{j-1}} L_j^{j-1} + \frac{\left( x_2 + \sum_{l=2}^{j} x_l \right) x_j^{j-3}}{2(j-2)! D_{j-1}} L_j^{j-2} \right], \tag{2.20}$$

where we have defined

$$D_k = \prod_{j=1}^{k} x_j. \tag{2.21}$$

The base case for induction is the 3rd sum, Eq. (2.19). We assume our hypothesis holds for all $j$ up to $k$ and show it holds for $j = k+1$. We start with

$$\sum_{n_N=0}^{L_N} \cdots \sum_{n_{k+1}=0}^{L_{k+1}} \left[ \frac{x_k^{k-2}}{(k-1)! D_{k-1}} L_k^{k-1} + \frac{\left( x_2 + \sum_{l=2}^{k} x_l \right) x_k^{k-3}}{2(k-2)! D_{k-1}} L_k^{k-2} \right] \tag{2.22}$$

and replace $L_k$ with $\frac{x_{k+1}}{x_k}(L_{k+1} - n_{k+1})$ to get

$$= \sum_{n_N=0}^{L_N} \cdots \sum_{n_{k+1}=0}^{L_{k+1}} \left[ \frac{x_k^{k-2}}{(k-1)! D_{k-1}} \left( \frac{x_{k+1}}{x_k} \right)^{k-1} (L_{k+1} - n_{k+1})^{k-1} + \frac{\left( x_2 + \sum_{l=2}^{k} x_l \right) x_k^{k-3}}{2(k-2)! D_{k-1}} \left( \frac{x_{k+1}}{x_k} \right)^{k-2} (L_{k+1} - n_{k+1})^{k-2} \right]$$
$$= \sum_{n_N=0}^{L_N} \cdots \sum_{n_{k+1}=0}^{L_{k+1}} \left[ \frac{x_{k+1}^{k-1}}{(k-1)! D_k} (L_{k+1} - n_{k+1})^{k-1} \right.$$

$$+\frac{\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{2(k-2)!D_k}\left(L_{k+1}-n_{k+1}\right)^{k-2}\Bigg].$$

In order to evaluate the $(k+1)$th sum, we expand the polynomials in $L_{k+1}$. We use the binomial theorem:

$$\begin{aligned}
= \sum_{n_N=0}^{L_N}\cdots\sum_{n_{k+1}=0}^{L_{k+1}} & \Bigg[\frac{x_{k+1}^{k-1}}{(k-1)!D_k}\sum_{i=0}^{k-1}(-1)^{k-1-i}\binom{k-1}{i}L_{k+1}^{i}n_{k+1}^{k-1-i} \\
& +\frac{\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{2(k-2)!D_k}\sum_{i=0}^{k-2}(-1)^{k-2-i}\binom{k-2}{i}L_{k+1}^{i}n_{k+1}^{k-2-i}\Bigg].
\end{aligned}$$

The coefficients of the polynomials in $L_{k+1}$ do not depend on $n_{k+1}$, so we pull those out and focus on the $(k+1)$th sum over the sums of the binomials. Expanding out the binomial sum, using the approximation from Lemma 2.2.1, and combining like terms, we find that

$$\begin{aligned}
\sum_{n_{k+1}=0}^{L_{k+1}}\sum_{i=0}^{k-1}(-1)^{k-1-i}\binom{k-1}{i}L_{k+1}^{i}n_{k+1}^{k-1-i} &\approx \frac{1}{k}L_{k+1}^{k}+\frac{1}{2}L_{k+1}^{k-1} \\
\sum_{n_{k+1}=0}^{L_{k+1}}\sum_{i=0}^{k-2}(-1)^{k-2-i}\binom{k-2}{i}L_{k+1}^{i}n_{k+1}^{k-2-i} &\approx \frac{1}{k-1}L_{k+1}^{k-1}+\frac{1}{2}L_{k+1}^{k-2}. \quad (2.23)
\end{aligned}$$

Plugging those results back into the sum gives

$$\begin{aligned}
= \sum_{n_N=0}^{L_N}\cdots\sum_{n_{k+2}=0}^{L_{k+2}} & \Bigg[\frac{x_{k+1}^{k-1}}{(k-1)!D_k}\left(\frac{1}{k}L_{k+1}^{k}+\frac{1}{2}L_{k+1}^{k-1}\right) \\
& +\frac{\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{2(k-2)!D_k}\left(\frac{1}{k-1}L_{k+1}^{k-1}+\frac{1}{2}L_{k+1}^{k-2}\right)\Bigg] \\
= \sum_{n_N=0}^{L_N}\cdots\sum_{n_{k+2}=0}^{L_{k+2}} & \Bigg[\frac{x_{k+1}^{k-1}}{k!D_k}L_{k+1}^{k}+\frac{x_{k+1}^{k-1}}{2(k-1)!D_k}L_{k+1}^{k-1} \\
& +\frac{\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{2(k-1)!D_k}L_{k+1}^{k-1}+\frac{\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{4(k-2)!D_k}L_{k+1}^{k-2}\Bigg] \\
= \sum_{n_N=0}^{L_N}\cdots\sum_{n_{k+2}=0}^{L_{k+2}} & \Bigg[\frac{x_{k+1}^{k-1}}{k!D_k}L_{k+1}^{k}+\frac{x_{k+1}^{k-1}+\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{2(k-1)!D_k}L_{k+1}^{k-1}
\end{aligned}$$

$$+\frac{\left(x_2+\sum_{l=2}^{k}x_l\right)x_{k+1}^{k-2}}{4(k-2)!D_k}L_{k+1}^{k-2}\Bigg].$$

We drop all terms of order $L_{k+1}^{k-2}$, leaving us with

$$\sum_{n_N=0}^{L_N}\cdots\sum_{n_{k+2}=0}^{L_{k+2}}\frac{x_{k+1}^{k-1}}{k!D_k}L_{k+1}^{k}+\frac{\left(x_2+\sum_{l=2}^{k+1}x_l\right)x_{k+1}^{k-2}}{2(k-1)!D_k}L_{k+1}^{k-1}. \tag{2.24}$$

This is the claimed form for the $(k+1)$th sum. By induction, our form holds for all $k=\{2,3,\ldots,N\}$. Hence, after the $N$th sum, our expression for $P_H(X)$ is

$$\frac{x_N^{N-2}}{(N-1)!D_{N-1}}L_N^{N-1}+\frac{\left(x_2+\sum_{l=2}^{N}x_l\right)x_N^{N-3}}{2(N-2)!D_{N-1}}L_N^{N-2}. \tag{2.25}$$

Recall that $L_N=\frac{X}{x_N}$. We substitute:

$$\begin{aligned}P_H(X) &\approx \frac{x_N^{N-2}}{(N-1)!D_{N-1}}\left(\frac{X}{x_N}\right)^{N-1}+\frac{\left(x_2+\sum_{l=2}^{N}x_l\right)x_N^{N-3}}{2(N-2)!D_{N-1}}\left(\frac{X}{x_N}\right)^{N-2}\\ &\approx \frac{1}{(N-1)!D_{N-1}}\frac{X^{N-1}}{x_N}+\frac{\left(x_2+\sum_{l=2}^{N}x_l\right)}{2(N-2)!D_{N-1}}\frac{X^{N-2}}{x_N}\\ &\approx \frac{X^{N-1}}{(N-1)!D_N}+\frac{X^{N-2}}{2(N-2)!D_N}\left(x_2+\sum_{l=2}^{N}x_l\right).\end{aligned} \tag{2.26}$$

This is the closed-form expression for $P_H(X)$ we desire, to second-largest order in $X$. Given Eq. (2.26), we can calculate an approximate $\langle n_j\rangle$. For ease of calculation, assume $x_j|X$ and let $\gamma=\left(x_2+\sum_{l=2}^{N}x_l\right)$. From Eq. (2.12), algebraic manipulations of our expressions yield

$$\langle n_j\rangle=\frac{X}{Nx_j}\left(1+\gamma\frac{(N-1)}{2X}\right)^{-1}\left(1-\frac{x_j}{X}\right)^{N}\left[1+\frac{\gamma N}{2X}\left(1-\frac{x_j}{X}\right)^{-1}\right]. \tag{2.27}$$

See Appendix A for a complete derivation of this expression.

As stated previously, we are interested in the regimes in which this quantity approaches the power law distribution, if at all. We have fixed $N$ and the $x_j$, but we may allow $X$ to grow arbitrarily large. As $X\to\infty$, we have that $\frac{x_j}{X}\to 0$, $\frac{\gamma N}{2X}\to 0$, and $\gamma\frac{(N-1)}{2X}\to 0$. Hence we are left with, in the large $X$ limit,

$$\langle n_j\rangle\approx\frac{X}{Nx_j}. \tag{2.28}$$

This holds for all $X \gg \gamma N$. If we make the approximation that $\gamma \approx N x_N$ (this is an overshoot), then our limiting behavior is valid when $X \gg N^2 x_N$. Hence we have determined the regime in which a power law distribution is a valid approximation. For reference, Eq. (2.27) is plotted against the Lemons' power law over a continuous range of $x_j$ in Figure 2.1.

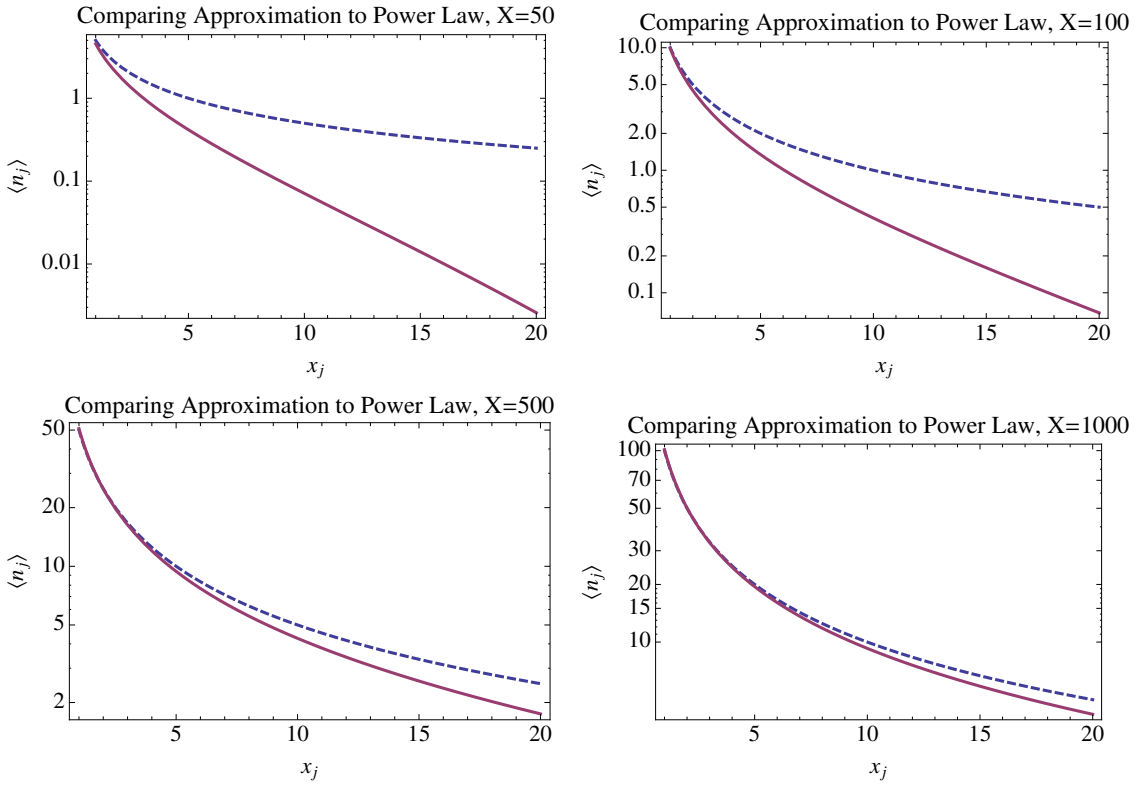



Figure 2.1: The approximation for $\langle n_j \rangle$ (solid line) is compared to the ideal power law Eq. (2.4) (dashed line) for $X = 50, 100, 500, 1000$. The range on our plots is on a log scale. Note that while the approximation and ideal are continuous over $x$, the exact $\langle n_j \rangle$ only has values at the integers.

### 2.2.3 Obtaining $\langle n_j \rangle$ by Bounding $P_H(X)$

Instead of choosing to approximate $P_H(X)$, we may instead derive closed-form bounds. Then we may use those to obtain upper and lower bounds for Eq. (2.12).

Consider again the sum expression for $P_H(X)$, Eq. (2.13), with $x_1 = 1$. As we evaluate sums, we find that our summands $f(n_j)$ are all positive and non-

increasing. Hence we may use the inequality [8]

$$\sum_{n=0}^{\lfloor L \rfloor} f(n) \geq \int_0^{\lfloor L \rfloor + 1} f(n) dn > \int_0^L f(n) dn, \tag{2.29}$$

where we have used the fact that $\lfloor L \rfloor > L - 1$. It follows, then, that

$$P_H(X) > \int_{n_N=0}^{L_N} \int_{n_{N-1}=0}^{L_{N-1}} ... \int_{n_2=0}^{L_2} dn_2 \cdots dn_N, \tag{2.30}$$

where the $L_j$ are defined by Eq. (2.18). It is fairly straightforward to integrate this expression, and the calculation is included in Appendix B. We arrive at a lower bound

$$P_H(X) > \frac{X^{N-1}}{(N-1)! D_N}. \tag{2.31}$$

For the upper bound, we again convert our sums into integrals. Here, however, we use the alternative inequality [8]

$$\sum_{n=0}^{\lfloor L \rfloor} f(n) \leq \int_{-1}^{\lfloor L \rfloor} f(n) dn \leq \int_{-1}^{L} f(n) dn = \int_0^{L+1} f(n' - 1) dn', \tag{2.32}$$

where $\lfloor L \rfloor \leq L$ and we have changed variables $n' = n + 1$. In terms of these variables, we note that

$$L_j = \frac{1}{x_j} \left( X + \sum_{k=j+1}^{N} x_k - \sum_{k=j+1}^{N} n'_k x_k \right). \tag{2.33}$$

We thus use one more inequality

$$L_j < L'_j = \frac{1}{x_j} \left( X' - \sum_{k=j+1}^{N} n'_k x_k \right), \tag{2.34}$$

where

$$X' = X + \sum_{k=2}^{N} x_k. \tag{2.35}$$

Altogether we find

$$P_H(X) < \int_{n'_N=0}^{L'_N} \int_{n'_{N-1}=0}^{L'_{N-1}} ... \int_{n'_2=0}^{L'_2} dn'_2 \cdots dn'_N. \tag{2.36}$$

These integrals can be evaluated as in the lower bound case to yield

$$P_H(X) < \frac{{X'}^{N-1}}{(N-1)!D_N} = \frac{1}{(N-1)!D_N}\left(X + \sum_{k=2}^{N} x_k\right)^{N-1}. \tag{2.37}$$

Thus we have bounded $P_H(X)$. Consider again Eq. (2.12). To get a lower bound for $\langle n_j \rangle$, we use the upper bound for $P_H(X)$ as our denominator and the lower bound for $P_H(X)$ in the sum (that is, we underestimate by dividing too small a number by too large a number). Using Eqs. (2.31) and (2.37), we get

$$\begin{aligned}
\langle n_j \rangle &> \frac{1}{\frac{{X'}^{N-1}}{(N-1)!D_N}} \sum_{i=1}^{\lfloor X/x_j \rfloor} \frac{(X - ix_j)^{N-1}}{(N-1)!D_N} \\
&= \frac{1}{{X'}^{N-1}} \sum_{i=1}^{\lfloor X/x_j \rfloor} (X - ix_j)^{N-1} \\
&> \frac{1}{{X'}^{N-1}} \int_1^{X/x_j} (X - zx_j)^{N-1} dz \\
&= \frac{1}{{X'}^{N-1}} \frac{1}{Nx_j} (X - x_j)^N \\
&> \frac{X}{Nx_j} \left(\frac{X - x_j}{X'}\right)^N \\
&> \frac{X}{Nx_j} \left(1 - \frac{x_j}{X} - \frac{1}{X}\sum_{k=2}^{N} x_k\right)^N,
\end{aligned} \tag{2.38}$$

where we have used the inequalities $(1+x)^{-1} > 1 - x$ and Eq. (2.29).

To get an upper bound for $\langle n_j \rangle$, we use the lower bound for $P_H(X)$ as our denominator and the upper bound for $P_H(X)$ in the sum (we overestimate by dividing too large a number by too small a number). Again using Eqs. (2.31) and (2.37), we have

$$\begin{aligned}
\langle n_j \rangle &< \frac{1}{\frac{X^{N-1}}{(N-1)!D_N}} \sum_{i=1}^{\lfloor X/x_j \rfloor} \frac{(X' - ix_j)^{N-1}}{(N-1)!D_N} \\
&= \frac{1}{X^{N-1}} \sum_{i=1}^{\lfloor X/x_j \rfloor} (X' - ix_j)^{N-1} \\
&< \frac{1}{X^{N-1}} \int_0^{X/x_j+1} (X' + x_j - zx_j)^{N-1} \, dz
\end{aligned}$$

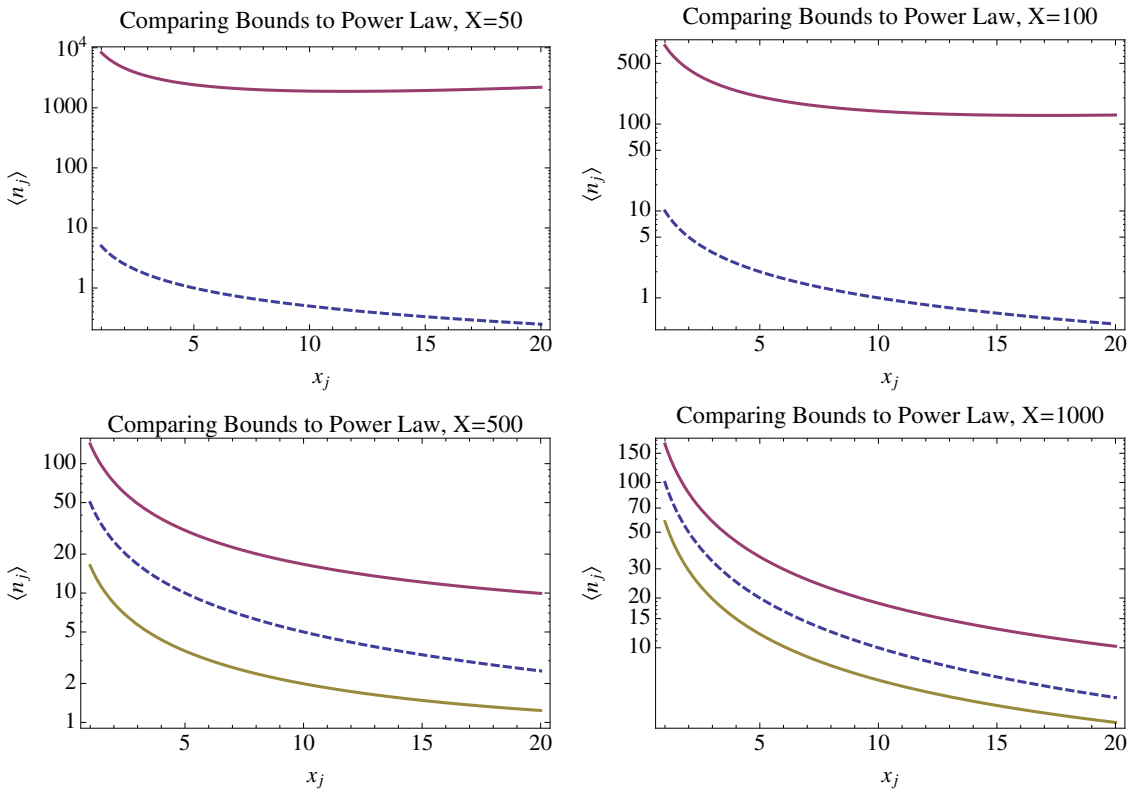



Figure 2.2: The bounds for $\langle n_j \rangle$ (solid lines) are compared to Eq. (2.4) (dashed line) for $X = 50, 100, 500, 1000$. The range on our plots is on a log scale. Note that while the bounds and ideal power law are continuous over $x$, the exact $\langle n_j \rangle$ only has values at the integers. For $X = 50$ and $X = 100$, the lower bound is order of magnitudes smaller than the power law, so it has been omitted from those plots.

$$\begin{aligned} &< \frac{1}{X^{N-1}} \frac{1}{N x_j} \left(X' + x_j\right)^N \\ &< \frac{X}{N x_j} \left(\frac{X' + x_j}{X}\right)^N \\ &< \frac{X}{N x_j} \left(1 + \frac{x_j}{X} + \frac{1}{X} \sum_{k=2}^{N} x_k\right)^N , \end{aligned} \tag{2.39}$$

where here we have used the inequality of Eq. (2.32). Taking Eqs. (2.38) and

(2.39) together, we conclude that

$$\frac{X}{Nx_j}\left(1-\frac{x_j}{X}-\frac{1}{X}\sum_{k=2}^{N}x_k\right)^N<\langle n_j\rangle<\frac{X}{Nx_j}\left(1+\frac{x_j}{X}+\frac{1}{X}\sum_{k=2}^{N}x_k\right)^N. \quad (2.40)$$

As in the approximation case, we must consider in which regime of $X$ these bounds converge to a power law behavior. We fix $N$ and all the $x_j$ and let $X\to\infty$. We have that $\frac{x_j}{X}\to 0$ and $\frac{Nx_N}{X}\to 0$. In the large $X$ limit, Eq. (2.40) becomes

$$\frac{X}{Nx_j}\left(1-\frac{Nx_j}{X}-\frac{N}{X}\sum_{k=2}^{N}x_k+\cdots\right)<\langle n_j\rangle<\frac{X}{Nx_j}\left(1+\frac{Nx_j}{X}+\frac{N}{X}\sum_{k=2}^{N}x_k+\cdots\right),$$

or

$$\langle n_j\rangle=\frac{X}{Nx_j}\left[1+\mathcal{O}\left(\frac{N^2x_N}{X}\right)\right]. \quad (2.41)$$

where $\mathcal{O}(x)$ indicates that there are additional terms in the expression of size at most a constant times $x$. Note that in the previous expression we had the expression $x_j+\sum_{k=2}^{N}x_k$, multiplied by the quantity $\frac{N}{X}$. Because $x_k\le x_N$ for all $k$, we know that

$$\begin{aligned}\frac{N}{X}\left(x_j+\sum_{k=2}^{N}x_k\right) &\le \frac{N}{X}\left(x_N+\sum_{k=2}^{N}x_N\right)\\ &= \frac{N}{X}(Nx_N)\\ &= \frac{N^2x_N}{X}. \end{aligned} \quad (2.42)$$

Hence we see that the magnitude of the largest nonunity terms in our expansion of Eq. (2.40) is bound above by $\frac{N^2x_N}{X}$. Thus the terms in the expansion will be at most a constant times this upper bound. Using that fact, we have rigorously proven that whenever $X\gg N^2x_N$, Lemons' result holds, as the extra terms tend to 0. This is in accordance with our result from the approximation method. As in the previous section, we include plots of Eq. (2.40) against the Lemons power law over a continuous range of $x_j$. See Figure 2.2.

### 2.2.4 The Canonical Ensemble

Rather than construct a new statistical model, we may instead cast our system as a previously established problem. Consider the quantum harmonic oscillator as presented in a typical undergraduate statistical physics course [17]. We establish

a total energy of $X$ energy quanta $\hbar\omega_0$ for our system. The energy may then be distributed over a set of quantum harmonic oscillators, where each oscillator can only accept a particular integer multiple of quanta. That is, oscillator $j$ may only be excited by multiples of $j\hbar\omega_0$ quanta. We may determine the average number $n_j$ of quanta in each oscillator by treating the $n_j$ as random variables with the distribution $\Pr(n_j) = A_j e^{-\beta x_j n_j}$, where $A_j$ is some normalization constant, $x_j$ corresponds to the energy factor associated with the oscillator $n_j$, and $\beta = (k_B T)^{-1}$ is our functional "temperature." Then

$$\langle n_j \rangle = \frac{\sum_{n=0}^{\infty} n e^{-\beta x_j n}}{\sum_{n=0}^{\infty} e^{-\beta x_j n}} = \frac{1}{e^{\beta x_j} - 1}. \tag{2.43}$$

Now, we establish the connection to our partition problem. Let each harmonic oscillator correspond to a piece size via the energy factor $x_j$. Then because $n_j$ keeps track of how many quanta of size $x_j$ are in the $j$th oscillator, we have that it counts the number of pieces of a given size. The total energy $X$ becomes the partitioned conserved quantity. Only $\beta$ remains without an analog, but we will find an expression for it in terms of known variables. We require that on average the total is conserved: $\sum_j \langle n_j \rangle x_j = X$. Thus, we we wish to solve

$$\sum_j \frac{x_j}{e^{\beta x_j} - 1} = X \tag{2.44}$$

for $\beta$. From Lemons and our previous results, we expect $\langle n_j \rangle$ to roughly correspond to $\frac{X}{N x_j}$, where $N$ again is the number of possible piece sizes. If $X$ may be arbitrarily large and $x_j$ fixed, then the denominator of Eq. (2.44) must be small. It follows that we must have $\beta$ small. This is plausible, as we would expect the high-temperature limit of statistical physics to correspond to higher energies $X$. With $\beta$ small, we can expand our denominator to second order:

$$e^{\beta x_j} - 1 \approx 1 + \beta x_j + \frac{1}{2}\beta^2 x_j^2 - 1 = \beta x_j \left(1 + \frac{1}{2}\beta x_j\right). \tag{2.45}$$

Then $\sum_j \langle n_j \rangle x_j = \frac{1}{\beta} \sum_j \left(1 + \frac{1}{2}\beta x_j\right)^{-1}$, which we approximate again, using a binomial expansion to first order:

$$\sum_j \langle n_j \rangle x_j \approx \frac{1}{\beta} \sum_j \left(1 - \frac{1}{2}\beta x_j\right) = N\left(\frac{1}{\beta} - \frac{\langle x_j \rangle}{2}\right), \tag{2.46}$$

where we have defined $\sum_j \frac{x_j}{N} = \langle x_j \rangle$. Setting Eq. (2.46) equal to $X$, we find that

$$\frac{1}{\beta} = \frac{X}{N} + \frac{\langle x_j \rangle}{2}. \tag{2.47}$$

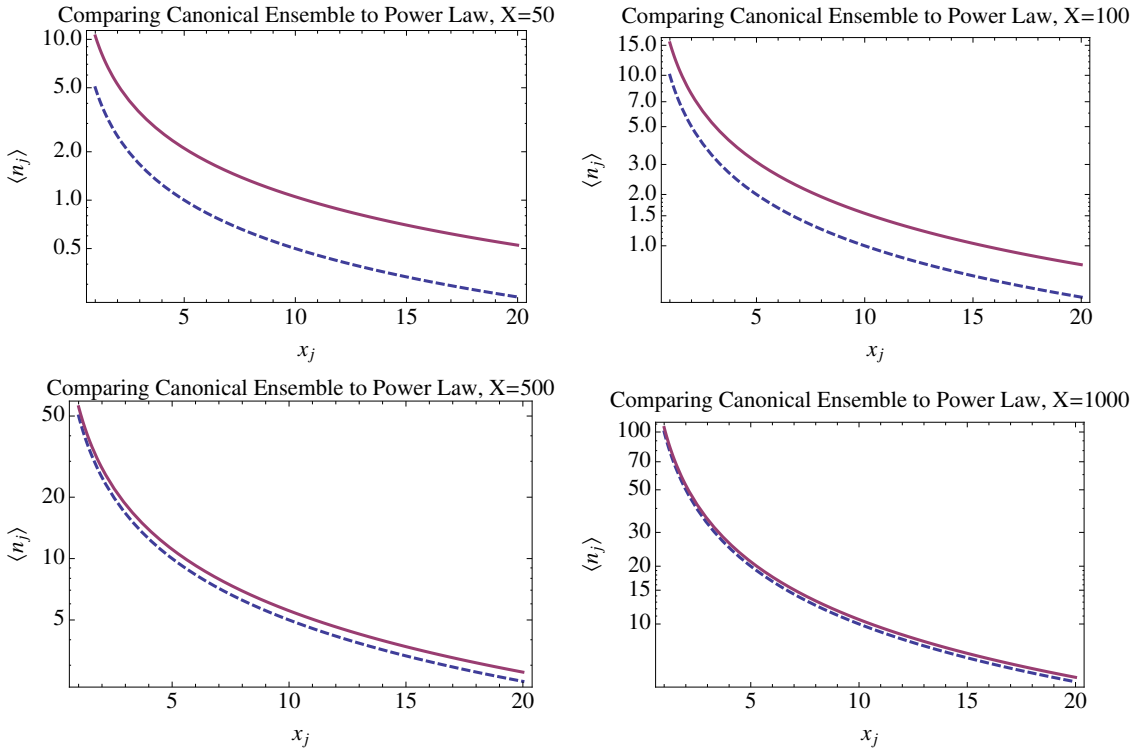



Figure 2.3: Eq. (2.48), the canonical ensemble model for $\langle n_j \rangle$ (solid line) is compared to Eq. (2.4) (dashed line) for $X = 50$, 100, 500,1000. The range on our plots is on a log scale. Note that while the bounds and ideal are continuous over $x$, the exact $\langle n_j \rangle$ only has values at the integers.

We substitute this expression into Eq. (2.43) and find that it becomes

$$\langle n_j \rangle = \frac{X}{N x_j} \left( 1 + \frac{N \langle x_j \rangle}{X} \right). \tag{2.48}$$

We have included a plot of Eq. (2.48) against the Lemons result for continuous $x_j$ in Figure 2.3 We may safely replace $\langle x_j \rangle$ with $x_N$, in which case we only require $X \gg N x_N$ in order for this expression to match that of Lemons. We note that this limiting regime differs from that of the previous approaches by a factor of $N$, suggesting that our approximation and bounds can be refined to hold for a smaller regime of $X$. In the following section we will compare the different models and approaches to determine which best describes our system.

### 2.2.5 Comparing the Approaches

In the preceding sections, we presented two models for the average number of pieces $n_j$ of size $x_j$. We wish to compare the results of the models against exact results. That is, we have used computational resources to enumerate all restricted partitions and numerically calculate the average number of pieces for a fixed set $H = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ over various $X$. This set was chosen for its simplicity. The following plots required an exact calculation of $P_H(X)$ for each $X$. We present a sampling of these results in Figure 2.4.

We first observe that as $X$ is increased, the model data points approach the calculated data points. Without any significant statistical analysis, one can see that the approximation method and canonical ensemble model exhibit a relatively similar relative error, though the former consistently undershoots and the latter consistently overshoots. Even so, those data points are certainly of the same order of magnitude as the calculated values. In stark contrast is the bounding approach. Both the upper and lower bounds are orders of magnitude off of the exact values for $X$ through 400. This suggests that while the bound approach should converge in the same $X$ regime as the approximation approach, it does noticeably worse for smaller $X$. Based on the relative success of the other methods, we believe that our bounds can be made significantly tighter. Specifically, the canonical ensemble result suggests that we can improve our statistical model in such a way that our approximation and bound methods converge to the power law for $X \gg Nx_N$ rather than the current $X \gg N^2 x_N$.

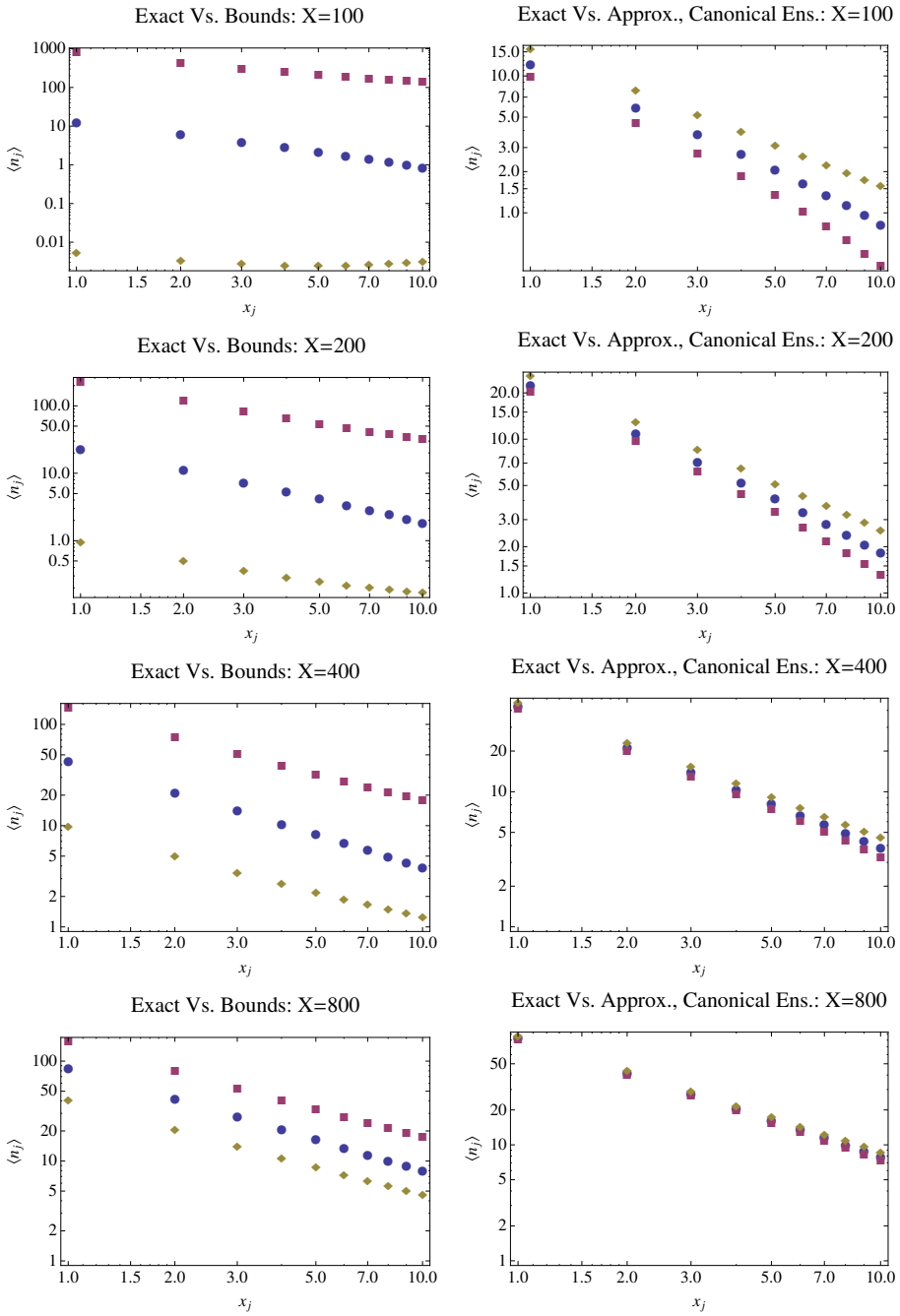



Figure 2.4: We plot the average number of pieces of size $x_j$ for our three models/methods alongside our exact values, calculated for various $X$. In each plot, the circular marks indicate the exact $\langle n_j \rangle$. On the left, the squares and diamonds indicate the upper and lower bounds, respectively. On the right, the squares and diamonds indicate the approximation and canonical ensemble model, respectively. The plots are on a log-log scale.

## 2.3 Variable Set Size

As one might expect, a power law distribution does not by itself guarantee Benford behavior, though it is a powerful indicator. An obvious example would be the set $H = \{1, 2, 3\}$. That is a valid set, but even following a power law distribution, it cannot be Benford, as the probabilities for most digits will be 0. So we must have a set featuring all ten digits. Even so, that is not necessarily enough.

Recall that Lemons took his bin sizes to 0, the continuum limit, so he could integrate the power law. But as we stated previously, the measurable systems to which we hope to apply our findings are not continuous, but rather have some finite base unit. So not only do we need all first digits present, we need the gradations between piece sizes to be sufficiently smaller than the entire domain of piece sizes. In this section we pursue a numerical model to test how fine that unit must be in order for a power law distribution to yield Benford digit probabilities. We will use a simple $H$ that contains all integers up to some specified integer $n$. Though we lose a great deal of generality, this is a fairly reasonable assumption and should describe a good number of systems we would like to model.

For a generic probability distribution, the likelihood of picking a piece at random whose first digit is $m$ is given by

$$\Pr(\text{first digit } m) = \sum_m I_m(x_j)\langle n_j\rangle \;/\; \sum_j \langle n_j\rangle, \tag{2.49}$$

where we define $I_m(x_j) \equiv 1$ if the first digit of $x_j$ is $m$ and $I_m(x_j) \equiv 0$ otherwise. This function serves to pick out only those $x_j$ that start with the digit $m$.

We consider the Lemons power law Eq. (2.4) over the set of piece sizes $H_N = \{1, 2, 3, \ldots, N\}$. We do not consider $X$ except to say that it is sufficiently large to ensure said power law. Eq. (2.49) becomes

$$\Pr(\text{first digit } m) = \sum_{x=1}^{N} \frac{I_m(x_j)}{x} \;/\; \sum_{x=1}^{N} \frac{1}{x}. \tag{2.50}$$

We can use computational software to compute this probability for all 9 digits over various $N$. We are interested in when a power law results in sufficiently Benford behavior in each digit. By this we mean that, suitably averaged, the relative error between the probability of a given first digit and the Benford probability is $\leq 10\%$. When we are dealing with large $n$, $|H_n|$ and $|H_{n+1}|$ differ by a negligible amount. Hence, we are more interested in the order of magnitude of $n$ than the exact value. We will compute Pr(first digit $m$) for an entire order of magnitude and take the relative error from Benford. Then we average those errors over the entire decade to obtain the average error. This is the value we cite in Table 2.1.

| $\eta$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|
| $m = 1$ | 0.2810 | 0.1792 | 0.1321 | 0.1047 | 0.0868 | 0.0741 | 0.0646 | 0.0573 |
| $m = 2$ | 0.1177 | 0.0744 | 0.0550 | 0.0436 | 0.0362 | 0.0309 | 0.0270 | 0.0239 |
| $m = 3$ | 0.1049 | 0.0654 | 0.0479 | 0.0378 | 0.0313 | 0.0266 | 0.0232 | 0.0206 |
| $m = 4$ | 0.1364 | 0.0855 | 0.0626 | 0.0494 | 0.0408 | 0.0348 | 0.0303 | 0.0269 |
| $m = 5$ | 0.1856 | 0.1177 | 0.0866 | 0.0685 | 0.0567 | 0.0484 | 0.0422 | 0.0374 |
| $m = 6$ | 0.2448 | 0.1569 | 0.1160 | 0.0921 | 0.0763 | 0.0652 | 0.0569 | 0.0505 |
| $m = 7$ | 0.3052 | 0.1973 | 0.1464 | 0.1165 | 0.0967 | 0.0827 | 0.0722 | 0.0641 |
| $m = 8$ | 0.3625 | 0.2359 | 0.1756 | 0.1400 | 0.1164 | 0.0996 | 0.0871 | 0.0773 |
| $m = 9$ | 0.4175 | 0.2732 | 0.2039 | 0.1629 | 0.1356 | 0.1161 | 0.1016 | 0.0902 |

Table 2.1: Relative error between calculated first-digit probabilities and the Benford probability $\log_{10}\left(1 + \frac{1}{m}\right)$. Here $\eta$ denotes the decade over which the relative orders were averaged, spanning $(10^{\eta-1} + 1)$ to $10^{\eta}$, inclusive.

For clarification, the quoted average for order $\eta$ is the average of the relative errors for all integers from $10^{\eta-1} + 1$ to $10^{\eta}$, inclusive. Values corresponding to $\eta = 1$ have been omitted from the table. We see for all $m$ that we are within 20% by order 4 and 10% by order 9. This is sufficient for our purposes. One may note that the percent errors seem to be loosely dependent on the digit $m$, in the sense that the larger digits usually have larger errors for each corresponding $\eta$.

What we are really interested in is the rate at which each digit approaches its Benford ideal. The numbers overall tell us that if $N \sim 10^9$, a power law distribution should give Benford first-digit probabilities within 10%. But we may want to know if one digit becomes Benford earlier than its compatriots. One simple way to quantify this is to normalize each percent error in such a way that we can determine by how much the error is reduced for each successive $\eta$. For each $m$, we divide the percent errors by the percent error for $\eta = 2$. Then we can tell by what factor the percent error changes for each increase in $\eta$. We compare these factors for all $m$. This analysis is included as Figure 2.5.

We see that the percent errors decrease linearly in our log-log scale. This is indicative of a power law (in $1/\eta$) decrease in percent error, decreasing at the same rate for all $m$. This is an agreeable conclusion, as each decade should contain the same number of terms with each potential starting digit. Thus we would expect each decade, adding ten times as many terms, should decrease the error in a similar way. Also, because Benford's Law is at its heart logarithmic, it is not surprising to see $1/\eta$ decrease in the error. Moreover, we could fit these lines to obtain the decay equation and to determine the Benford error for any given $\eta$.

This scheme may have its limitations, but in conjunction with the previous sections, it gives us a sense of how large a system we must deal with in order to

see Benford behavior emerge.

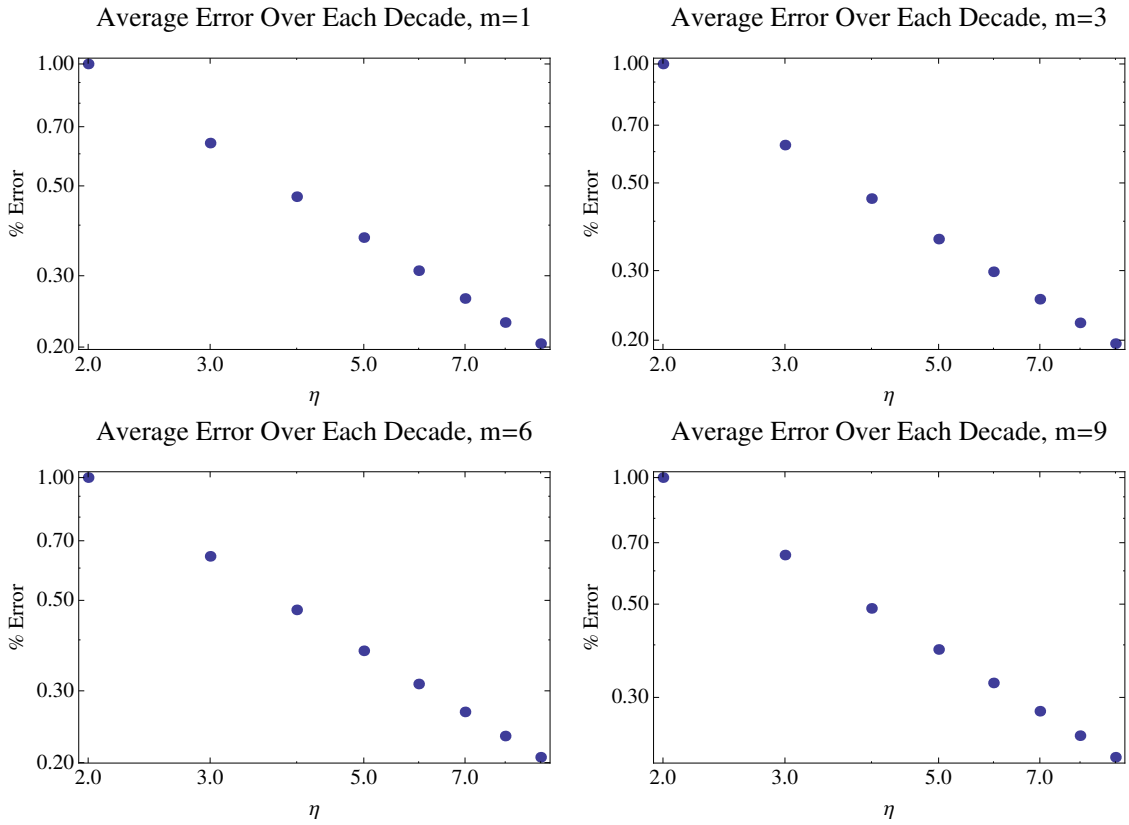



Figure 2.5: Normalized relative error for $m = 1, 3, 6, 9$. The scale is log-log. Notice that the plots are nearly identical, and all linear.

# Chapter 3

# Fragmentation Processes

## 3.1 A Stochastic Approach

In the previous chapter, we analyzed the aftermath of a generalized fragmentation process. That is, our process was essentially a black box that, given a conserved quantity, produced fragment distributions, and as we often do in statistical physics, we assumed all outcomes were equally likely. However, it is often the case that the process itself plays a role in determining the likelihood of various outcomes. For instance, it may be that in a certain procedure, pieces can recombine with other pieces after breaking, resulting in larger resulting pieces. Or instead, physical considerations may render all pieces larger than the base piece size unstable. Thus it is useful to consider classes of fragmentation processes and the resulting distributions.

Furthermore, we wish to consider the role of time. We could characterize a system at any moment in a process, but we want to know the long-term distribution. Specifically, we are interested in systems undergoing constant fragmentation that eventually reach a steady ratio of pieces. We call these steady states. Our goal is to characterize a process by the piece distribution of its steady state.

In what follows, we examine a few general classes of fragmentation processes and determine the steady state behavior of each. We note under which conditions, if any, the steady state follows a power law distribution. These then are the processes under which we might expect Benford behavior, given the range of piece sizes is sufficiently large (see Section 2.3 for an extended discussion). For numerical tests of some specific fragmentation processes for Benford behavior, see [4].

### 3.1.1 The Master Equation

We will model our fragmentation processes as a matrix differential equation. First, we must have some matrix $\mathbf{M}$ that governs how pieces of different sizes relate to one another. We call $\mathbf{M}$ the *transition matrix* for our process. This terminology is reminiscent of time-independent Markov chains [19], in which the elements of the transition matrix give the probability of transitioning from one state of the system to another. Going off that analogy, we observe that $\mathbf{M}$ must be a square $m \times m$ matrix, where $m$ is the number of possible piece sizes (states of our system). Then the matrix element $M_{jk}$ denotes how pieces of size $k$ are broken into parts of size $x_j$ (we will see later why this is not exactly a probability). Thus if we know the likelihood of pieces of a given size breaking or recombining into pieces of another size, we can construct a transition matrix $\mathbf{M}$ for our system.

Given $\mathbf{M}$, we can model our fragmentation as a matrix differential equation. That is, the rate at which our fragment distribution changes is given by the vector equation

$$\frac{d\vec{n}}{dt} = \mathbf{M}\vec{n}. \tag{3.1}$$

Alternatively, we can look at how the number of pieces $n_j$ of a particular size $x_j$ changes via

$$\frac{dn_j}{dt} = \sum_k \mathbf{M}_{jk} n_k, \tag{3.2}$$

where the sum is over all piece sizes. This is called the *master equation* [19].

Now, we are still interested in the total quantity being conserved, so we again impose the conservation relation Eq. (2.5) on our system. Taking the derivative of both sides of that relation with respect to time, we see that

$$\frac{d}{dt}(X) = 0 = \sum_j x_j \frac{dn_j}{dt} = \sum_{jk} x_j \mathbf{M}_{jk} n_k, \tag{3.3}$$

where the sum over $j$ is over all piece sizes.

We want the above to hold true for all values of $n_k$, so we require

$$\sum_j x_j \mathbf{M}_{jk} = 0. \tag{3.4}$$

We use the above method to construct all of our transition matrices.

### 3.1.2 Single-Path: Directed

For our first example, consider a system consisting of four possible piece sizes, denoted $x_j$ for $j = \{1, 2, 3, 4\}$, with $x_j > x_k$ if $j > k$ (we use this system as the

basis for a majority of the examples in this chapter). Each piece may break into pieces of the size directly smaller. So $x_4$ can break into $x_3$, $x_3$ can break into $x_2$, $x_2$ can break into $x_1$, and $x_1$ cannot break into any other piece sizes. We can represent this fragmentation procedure as a directed graph. A *directed graph* is a graph in which every edge is associated with a single direction [15]. So while it is possible to travel from one node to a connected node, the reverse transition cannot be made. Figure 3.1a shows our proposed fragmentation process as a directed graph. We call this a *single-path* directed graph because each piece may only break into one other piece size. That is, no more than one edge leads away from any node.

When a piece of size $x_4$ breaks up, what are we left with? We know we will have pieces of size $x_3$, but how many? Well, we require that the total quantity in the system be conserved, so such a break-up cannot add or take away from the system. That is, we need $n_4 x_4 = n_3 x_3$, so if $n_4 = 1$ for one piece breaking, we are left with $\frac{x_4}{x_3}$ pieces of size $x_3$. Note this may not be an integer number of pieces. The steady state in many cases will just be the average steady distribution.

So for every piece of size $x_4$ that our system loses, it gains $\frac{x_4}{x_3}$ pieces of size $x_3$. The case of breaking $x_3$ and $x_2$ follow similarly. Now, we assume that every break-up occurs with the same frequency or likelihood. Hence we can construct a transition matrix for our process:

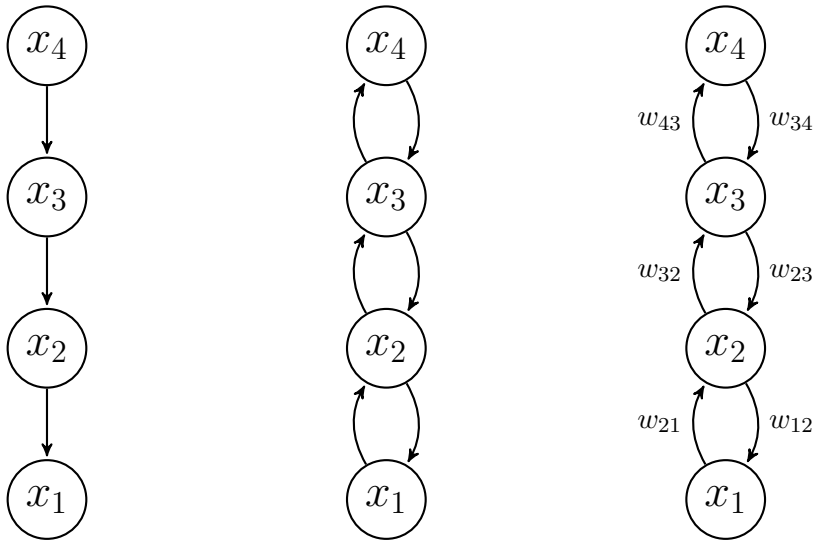
$$\mathbf{M_{sd}} = \begin{pmatrix} 0 & x_2/x_1 & 0 & 0 \\ 0 & -1 & x_3/x_2 & 0 \\ 0 & 0 & -1 & x_4/x_3 \\ 0 & 0 & 0 & -1 \end{pmatrix}. \tag{3.5}$$

The $-1$ along the diagonal indicate a piece breaking. The reader may confirm that Eq. (3.4) is satisfied, so $\mathbf{M_{sd}}$ is a valid transition matrix. One could take an initial piece distribution vector $\vec{x}_{int}$ and evaluate $\mathbf{M_{sd}}\vec{x}_{int}$ to find the piece distribution after each piece has been allowed to fragment.

Now, we want the steady state behavior of this system, so we have to solve the differential equation Eq. (3.1). We see that setting $\frac{d\vec{n}}{dt} = 0$ is equivalent to setting $\mathbf{M}\vec{n} = 0$. Thus, to find the steady state, we simply find the eigenvalues of $\mathbf{M_{sd}}$. If one such eigenvalue is 0, then there exists some vector $\vec{v}$ such that $\mathbf{M_{sd}}\vec{v} = 0\vec{v}$, and $\vec{v}$ is the desired steady state. The eigenvalue and eigenvector calculations are omitted here. We will just quote the steady state vectors.

Calculations show that 0 is indeed an eigenvalue of $\mathbf{M_{sd}}$, with corresponding eigenvector $\vec{n}_{sd} = (1, 0, 0, 0)^T$. Unless otherwise stated, we normalize our steady state vectors so that the first entry is 1. To find out how many pieces of a given size would be present given a starting conserved quantity $X$, we sum the $n_j x_j$ and multiply by a constant factor $A$ so that the conservation relation is satisfied. Then the number of pieces of size $x_j$ is the $j$th element of $A\vec{n}_{sd}$. This steady

state does not follow a power law distribution, as the only remaining pieces are those of size $x_1$. This result follows intuitively, as we would expect a system that can only fragment into smaller pieces to eventually be exclusively in the smallest pieces.




(a) Directed (b) Undirected (c) Undirected, Weighted

Figure 3.1: Graph representations of single-path fragmentation processes.

### 3.1.3 Single-Path: Undirected

We now consider a system in which smaller pieces can combine together into pieces of the next largest size. So in the example above, pieces of size $x_3$ can now combine into pieces of size $x_4$, and so on. However, pieces may only transition into pieces the next size larger or the next size smaller. The graph representation of this process is given in Figure 3.1b. Note that this is very similar to Figure 3.1a. The only change is that each of the edges now goes in both directions (represented as two edges in the figure). We call this type of graph an *undirected graph* [15].

This changes our transition matrix. The elements corresponding to larger pieces breaking into smaller remain unchanged. But now we must consider: how many pieces of size $x_4$ do we get when a piece of $x_3$ recombines? Of course, a single piece cannot recombine into something larger than itself. The elements of our matrix simply keep track of how much we gain, on average, when we exchange one piece size for another. So we are content with a fractional answer here. If needed, we can adjust the result to 1 piece of size $x_4$ to see how many pieces of $x_3$

we might need. Earlier we determined that $x_4/x_3$ pieces of size $x_3$ resulted from one $x_4$. Inverting that result gives that one $x_3$ begets $x_3/x_4$ pieces of size $x_4$. By definition, this ratio is less than 1.

Continuing with this for the other piece sizes, our new transition matrix becomes

$$\mathbf{M_{su}} = \begin{pmatrix} -1 & x_2/x_1 & 0 & 0 \\ x_1/x_2 & -2 & x_3/x_2 & 0 \\ 0 & x_2/x_3 & -2 & x_4/x_3 \\ 0 & 0 & x_3/x_4 & -1 \end{pmatrix}. \tag{3.6}$$

See that $(\mathbf{M_{su}})_{22} = (\mathbf{M_{su}})_{33} = -2$ rather than $-1$ as before. This is because at each step we lose a piece to breaking and to recombining (though again, it really takes more than one piece to recombine). A quick calculation shows that this satisfies Eq. (3.4).

With this matrix in hand, we find that 0 is again one of its eigenvalues. This corresponds to the steady state eigenvector $\vec{n}_{su} = \left(1, \frac{x_1}{x_2}, \frac{x_1}{x_3}, \frac{x_1}{x_4}\right)^T$. If we write 1 as $\frac{x_1}{x_1}$, then we see that this vector is equivalent to $x_1 \left(\frac{1}{x_1}, \frac{1}{x_2}, \frac{1}{x_3}, \frac{1}{x_4}\right)^T$. That is, the $j$th entry is inversely proportional to $x_j$. This is exactly the power law behavior we seek. We conclude that the steady state of the single-path undirected fragmentation follows a power law distribution.

### 3.1.4 Single-Path: Weighted Undirected

The last two examples were predicated on all transitions being equally likely. Suppose instead that some transitions are more likely than others. Then we can assign to each transition from $x_k$ to $x_j$ a weight $w_{jk}$. So the rate at which $x_4$ break up into $x_3$ would be given the weight $w_{43}$. Of course, the elements along the diagonals will no longer depend only on the number of edges away from each node. Instead, the $jj$th element of our transition matrix is the additive inverse of the sum of the weights on the edges leading out the $j$th node. For instance, if we denote our transition matrix as $\mathbf{M_{wsu}}$, then $(\mathbf{M_{wsu}})_{33} = -(w_{23} + w_{43})$. That is, the rate at which we lose pieces of size $x_3$ depends on how quickly those pieces are breaking up into $x_2$ and how quickly they are combining into $x_4$. See Figure 3.1c for a graph representation.

Our transition matrix is a slightly altered version of Eq. (3.6):

$$\mathbf{M_{wsu}} = \begin{pmatrix} -w_{21} & w_{12}(x_2/x_1) & 0 & 0 \\ w_{21}(x_1/x_2) & -(w_{12} + w_{32}) & w_{23}(x_3/x_2) & 0 \\ 0 & w_{32}(x_2/x_3) & -(w_{23} + w_{43}) & w_{34}(x_4/x_3) \\ 0 & 0 & w_{43}(x_3/x_4) & -w_{34} \end{pmatrix}. \tag{3.7}$$

The weights are bothersome, but Eq. (3.4) is still satisfied. Thus, our quantity will be conserved regardless of the weight values.

That said, our steady state will depend on those weights. 0 is an eigenvalue of $\mathbf{M_{wsu}}$ corresponding to the steady state $\vec{n}_{swu} = \left(1, \frac{w_{21}}{w_{12}}\frac{x_1}{x_2}, \frac{w_{21}w_{32}}{w_{12}w_{23}}\frac{x_1}{x_3}, \frac{w_{21}w_{32}w_{43}}{w_{12}w_{23}w_{34}}\frac{x_1}{x_4}\right)^T$. It's fairly clear that certain values of the weights will not result in a power law distribution. But on the other hand, we can constrain the weights to elicit the desired behavior. Again we write 1 as $x_1/x_1$, so we can pull $x_1$ out of each element of the eigenvector. Then we need the remaining elements to be of the form $1/x_j$. That is only possible if each expression of weights simplifies to 1. That would require $w_{21} = w_{12}$. This then makes the third element coefficient $\frac{w_{21}w_{32}}{w_{12}w_{23}} = \frac{w_{32}}{w_{23}}$, so we must also have $w_{32} = w_{23}$. Going one step further, we conclude $w_{43} = w_{34}$. These equalities all say the same thing: the weight on each edge of the graph is the same in both directions. Put another way, we need $x_j$ to break into $x_k$ at the same rate at which $x_k$ recombine into $x_j$. Note that there is no required relation between the weights on different edges. So it might be much easier to go between $x_1$ and $x_2$ than between $x_2$ and $x_3$, but that comparison has no effect on the steady state.

### 3.1.5 Multi-Path

Consider a small child playing with a collection of LEGO® bricks. He has assembled a small stack of 10 bricks and handed it to you, claiming it is a spaceship. In an attempt to ruin his day, you throw the stack to the ground and it shatters. Can it only break into stacks of 9 bricks? No, it can shatter into any multiple of a single brick, up to 10. If we attempt to explain this fragmentation to the now hysterical toddler with the single-path processes above, we will find ourselves incapable. Instead, we need a model that can allow for a piece to break up into any allowable size smaller than itself.

We introduce the notion of the *multi-path* graph. We now allow for any number of edges to lead away from each node (hence the name). First, let all of the edges be equally weighted in each direction, and consider an edge between each pair of nodes in our system. See Figure 3.2. Our transition matrix will no longer contain any 0 elements, as each piece can now break or combine into any other piece. We simply extend $\mathbf{M_{su}}$, altering the diagonal to account for the fact that we must consider three transitions for each piece. Thus,

$$\mathbf{M_{mu}} = \begin{pmatrix} -3 & x_2/x_1 & x_3/x_1 & x_4/x_1 \\ x_1/x_2 & -3 & x_3/x_2 & x_4/x_2 \\ x_1/x_3 & x_2/x_3 & -3 & x_4/x_3 \\ x_1/x_4 & x_2/x_4 & x_3/x_4 & -3 \end{pmatrix}. \tag{3.8}$$

Indeed, Eq. (3.4) is satisfied.

Note $\mathbf{M_{mu}}$ has 0 as an eigenvalue, so it has a steady state corresponding to the eigenstate $\vec{n}_{mu} = \left(1, \frac{x_1}{x_2}, \frac{x_1}{x_3}, \frac{x_1}{x_4}\right)^T$, just as in the case of the single-path undirected. This implies that despite the extra paths, the undirected graphs have the same limiting behavior. In the analogy of travel, it may be quicker to get from point $A$ to $D$ if one can take a plane directly there. But if one has to fly first to $B$ and then to $C$ before traveling to $D$, that does not change the end result. The multi-path graph simply opens up more airports for our piece frequent flyers.

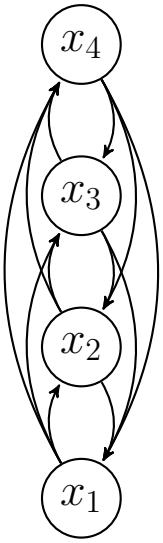



Figure 3.2: Graph representation of the multi-path fragmentation process. We may choose to weight the edges with weight $w_{jk}$, denoting the edge from $x_k$ to $x_j$.

Lastly, we wish to weight the edges of the multi-path undirected graph. We use the weighting convention of Section 3.1.4. The transition matrix is now

$$\mathbf{M_{wmu}}= \begin{pmatrix} -(w_{21}+w_{31}+w_{41}) & w_{12}(x_2/x_1) & w_{13}(x_3/x_1) & w_{14}(x_4/x_1) \\ w_{21}(x_1/x_2) & -(w_{12}+w_{32}+w_{42}) & w_{23}(x_3/x_2) & w_{24}(x_4/x_2) \\ w_{31}(x_1/x_3) & w_{32}(x_2/x_3) & -(w_{13}+w_{23}+w_{43}) & w_{34}(x_4/x_3) \\ w_{41}(x_1/x_4) & w_{42}(x_2/x_4) & w_{43}(x_3/x_4) & -(w_{14}+w_{24}+w_{34}) \end{pmatrix}. \quad (3.9)$$

As in all cases above, the quantity is conserved and Eq. (3.4) is still satisfied.

$\mathbf{M_{wmu}}$ has an eigenvalue of 0, but unfortunately the corresponding eigenvector is quite complicated and has been omitted in interest of space. We found that certain combinations of weights do lead to power law behavior. Testing different sets of edge weights, we found that the following system of equations must be satisfied to render the power law distribution:

$$
\begin{aligned}
w_{12} + w_{13} + w_{14} &= w_{21} + w_{31} + w_{41} \\
w_{21} + w_{23} + w_{24} &= w_{12} + w_{32} + w_{42} \\
w_{31} + w_{32} + w_{34} &= w_{13} + w_{23} + w_{43} \\
w_{41} + w_{42} + w_{43} &= w_{14} + w_{24} + w_{34}.
\end{aligned} \tag{3.10}
$$

These equations essentially say that the rate at which a piece is produced by other pieces must be equal to the rate at which that piece is fragmenting. This is just a generalization of the requirement in the weighted single-path undirected case. Note this model gives us much more flexibility than our previous attempts. We can set certain weights to 0 if we desire, should any transitions be disallowed in a system. We can tweak the weights here to get any of the previous examples.

Overall, we conclude that as long as the breaking and recombining processes in a system occur at the same rate, we should see power law behavior emerge. This is a useful benchmark when examining a physical fragmentation process.

## 3.2 Probing the Robustness of Power Law Behavior

In fragmentation processes that take the form of either a single-path or multi-path undirected graph, we have seen by example that a power law fragment distribution is attainable. Of course, this is only the case when the total rate at which $x_j$ breaks up is equal to the sum of the rates at which the other pieces combine into $x_j$, for all $j$. But what if the rates are just slightly out of line? Maybe fragmenting is slightly more preferred than recombining. We might expect the end distribution to be fairly close to a power law. But if we turned off recombination altogether, so that only fragmentation is possible, we would have a directed process. Then we would get a steady state distribution with only the smallest piece size. This leads to a natural question: as recombination becomes less likely in comparison to fragmentation, does the steady state distribution vary smoothly from a power law to a directed distribution? Or is there some critical midpoint at which the limiting behavior shifts from approximating one extreme to the other? This is the subject of this section. We first approach this problem numerically. Then, we make use of first-order perturbation theory to confirm our numerical results.

### 3.2.1 The Numerical Method

Let's formalize the question at hand. The single-path undirected graph can be thought of as the addition of two nearly-identical directed graphs (Figure 3.3).

The only difference between the two graphs is the direction of the edges. In the first, the edges all point to smaller piece sizes. We recognize this as the single-path directed graph. It deals exclusively with pieces breaking into smaller pieces. The other directed graph features all edges pointing to larger piece sizes. This graph deals exclusively with recombinations.

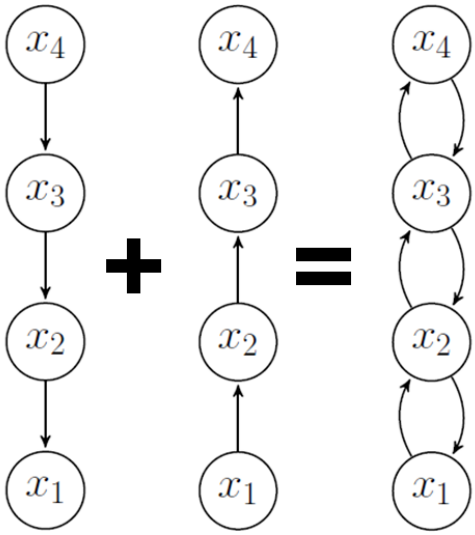



Figure 3.3: We see that a single-path undirected graph can be constructed from a graph dealing solely with fragmentations and a graph dealing solely with recombinations.

We can deconstruct the corresponding transition matrix in the same way. While we have no experience with a transition matrix corresponding to the recombination graph, we can use the intuition we have gained to write one, which we denote $\mathbf{M_{up}}$. Note also that it must add to the transition matrix for the directed graph, hereafter denoted $\mathbf{M_{down}}$, to yield the undirected transition matrix $\mathbf{M_0}$. That is, $\mathbf{M_0} = \mathbf{M_{down}} + \mathbf{M_{up}}$.

Now, this assumes an equal weighting between $\mathbf{M_{down}}$ and $\mathbf{M_{up}}$. Let $\delta \in [0, 1]$. We can multiply $\mathbf{M_{up}}$ by a weight $(1 - \delta)$ and examine the sum

$$\begin{aligned} \mathbf{M}(\delta) &= \mathbf{M_{down}} + (1-\delta)\mathbf{M_{up}} \\ &= (\mathbf{M_{down}} + \mathbf{M_{up}}) - \delta\mathbf{M_{up}} \\ &= \mathbf{M_0} - \delta\mathbf{M_{up}}. \end{aligned} \tag{3.11}$$

When $\delta = 0$, this sum is just the transition matrix for the undirected graph. Similarly, when $\delta = 1$, we have the transition matrix for the directed graph. Hence, by varying $\delta$ between 0 and 1, we can sweep from undirected to directed

behavior. If we replace the single-path transition matrices with their multi-path counterparts, the analysis is the same.

Now, we wish to calculate the steady state for a number of intermediate values of $\delta$. Then we can compare each limiting behavior to the two extremes. But to do so, we need a quantitative sense of how close the intermediate is to one or the other. Thus we introduce a function $F$ that takes a steady state as an input. We call this the *distribution estimator*.

We construct $F$ by starting with a sum-of-squares calculation comparing the input steady state to the power law distribution:

$$\sum_{j=1}^{N} \left| n_j^{(\delta)} - \frac{1}{Nx_j} \right|^2, \tag{3.12}$$

where $n_j^{(\delta)}$ is the $j$th element of the steady state vector corresponding to the perturbation $\delta$. Note that the power law has been normalized in such a way that the total quantity is 1. That is, $\sum n_j x_j = 1$. This makes $F$ more illuminating.

If we have a power law, this sum is 0. But if instead we have the directed graph, this sum evaluates to

$$\left| 1 - \frac{1}{Nx_1} \right|^2 + \sum_{j=2}^{N} \left| \frac{1}{Nx_j} \right|^2. \tag{3.13}$$

We would like to constrain $F$ to the range $[0, 1]$. Thus we divide Eq. (3.12) by Eq. (3.13) to get the distribution estimator

$$F(\delta) = \frac{\sum_{j=1}^{N} \left| n_j^{(\delta)} - \frac{1}{Nx_j} \right|^2}{\left| 1 - \frac{1}{Nx_1} \right|^2 + \sum_{j=2}^{N} \left| \frac{1}{Nx_j} \right|^2}. \tag{3.14}$$

When $F = 0$ we have a power law, and when $F = 1$ we have the directed case. The reader can see that any intermediate case will yield a $F$ value between those. We say that a steady state is closer to the power law distribution if $F < \frac{1}{2}$ and closer to the directed distribution if $F > \frac{1}{2}$. We treat $F = \frac{1}{2}$ as the turning point.

Given $F$, we can characterize the steady states for any $\delta$ in $[0, 1]$.

### 3.2.2 Results

We used *Mathematica* to generate perturbed transition matrices for $N = 10$, 50, 100, 500, 1000. For ease of calculation we considered cases where the piece sizes $x_j = j$. We incremented our perturbation parameter $\delta$ in multiples of $\frac{1}{100}$, so we considered 101 different perturbation scenarios. Given a $\delta$, *Mathematica*

computed the eigenvalues for $\mathbf{M}(\delta)$ and returned an error if 0 was not found among them. That was never the case, confirming that for every $\delta$ in the given range, a steady state exists. *Mathematica* then normalized the steady state and computed $F(\delta)$. These values were stored and are plotted for the single-path case in Figure 3.4, for various $N$.

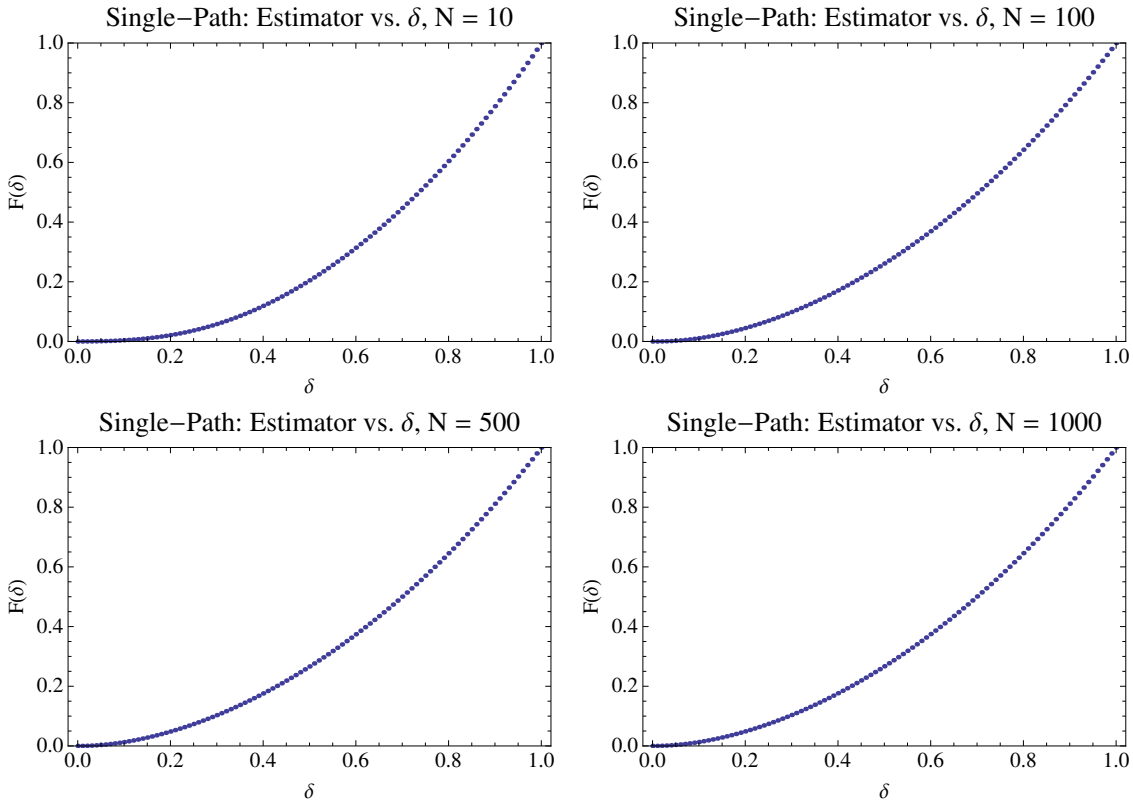



Figure 3.4: $F(\delta)$ plotted for various $N$ in the single-path case. The plots appear to be independent of $N$.

The striking feature of these plots is that they appear to be independent of $N$. That is, it seems $F$ is solely a function of $\delta$. Moreover, the transition from directed to power law distribution is smooth, though not linear. There is no abrupt change in behavior. The turning point $F = \frac{1}{2}$ occurs around $\delta \approx 0.75$, so the steady states are closer to a power law for a majority of perturbations.

Next we consider the multi-path case. We followed the same procedure as the single-path case in *Mathematica* and arrived at the plots in Figure 3.5.

These plots are striking for an altogether different reason. We see that the steady states are closer to power law for a majority of perturbations. Furthermore, as $N$ increases, $F$ remains closer to 0 for larger and larger $\delta$. Note that this leads

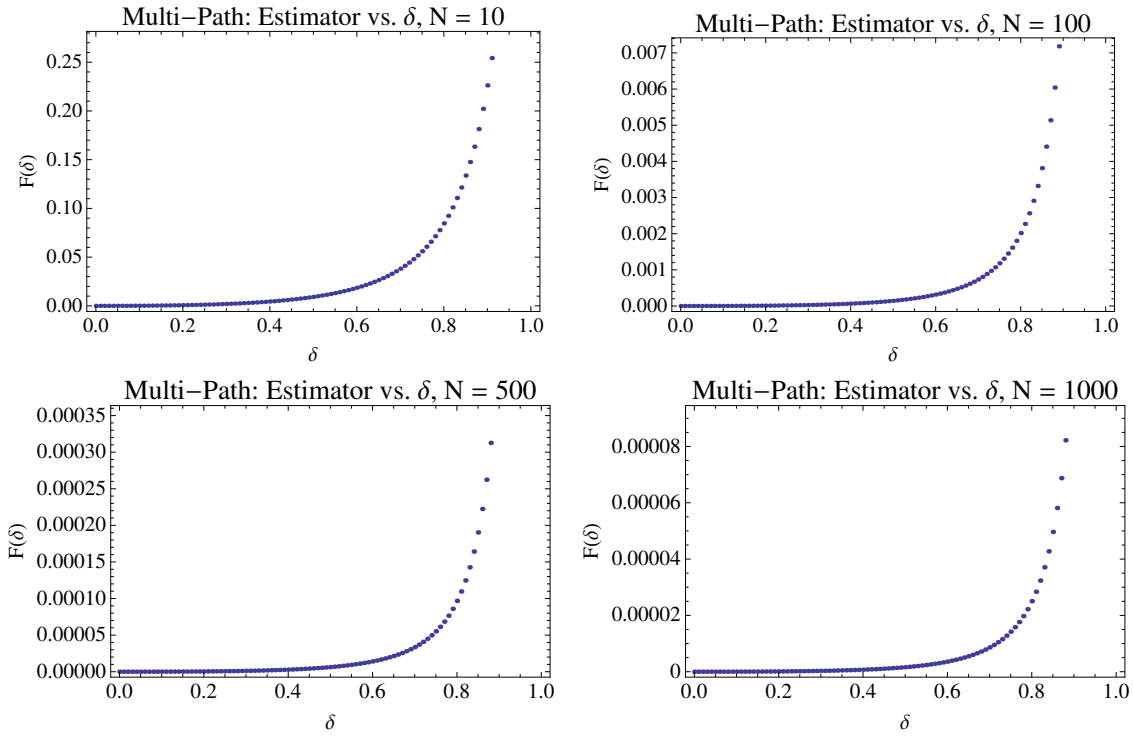



Figure 3.5: $F(\delta)$ plotted for various $N$ in the multi-path case. Note that each plot has a different vertical scale. As $N$ grows, $F$ stays closer to 0 for larger $\delta$.

to a wide jump in $F$ between $\delta = \frac{100}{101}$ and $\delta = 1$. We can get a better sense of how $F$ varies with $\delta$ by comparing the $F$ for various $N$ on one plot, Figure 3.6. We can also compute $\delta_{50}$, the perturbation parameter for which $F$ equals the turning point $\frac{1}{2}$. To do so, we have *Mathematica* interpolate a continuous function into our list of $F$ values. Then, we find the value of $\delta$ (now a continuous variable) for which $F = \frac{1}{2}$. These values are recorded in Table 3.1. See that $\delta_{50}$ is close to 1 even for small matrix sizes. By $N = 50$, $\delta_{50}$ is already above 99%. We would need perturbations finer than those we have considered in order to see that turning point. We project that as $N$ grows, $\delta_{50}$ continues to approach 1.

| $N$ | 10 | 50 | 100 | 500 | 1000 |
|---|---|---|---|---|---|
| $\delta_{50}$ | 0.96039 | 0.99153 | 0.99473 | 0.99654 | 0.99666 |

Table 3.1: $\delta_{50}$ is the projected value of the perturbation parameter for which $F = \frac{1}{2}$. We see that for $N$ as small as 50 this value is already above 0.99, indicating that in our system with perturbations in increments of $\frac{1}{100}$, the turning point occurs too close to the endpoint $\delta = 1$ to be observed.

This is indicative of a potential *phase transition*. A phase transition is an abrupt change in the properties of a system as an input parameter is varied infinitesimally [17]. That is, there is some critical parameter value $\delta_C$ (a function of $N$) such that for $\delta > \delta_C$, the system displays radically different behavior than for $\delta < \delta_C$. Based on the results, we expect that as $N \to \infty$, $F$ will become discontinuous at $\delta = 1$, a sudden change from power law behavior to directed.

We conclude that in the multi-path case, the power law distribution is very robust. As $N$ grows, it takes a larger and larger perturbation to appreciably disrupt the steady state. In the limit where $N$ grows arbitrarily large, we conclude that for any $\delta < 1$, we will see power law behavior. Only when recombination stops entirely will the power law vanish. Thus for any real-world fragmentation process that can be modeled by this multi-path fragmentation, we can be fairly confident that if the set of possible fragment sizes is sufficiently large, a perturbation will not unsettle the power law behavior, hence allowing us to conjecture whether that system will be Benford.

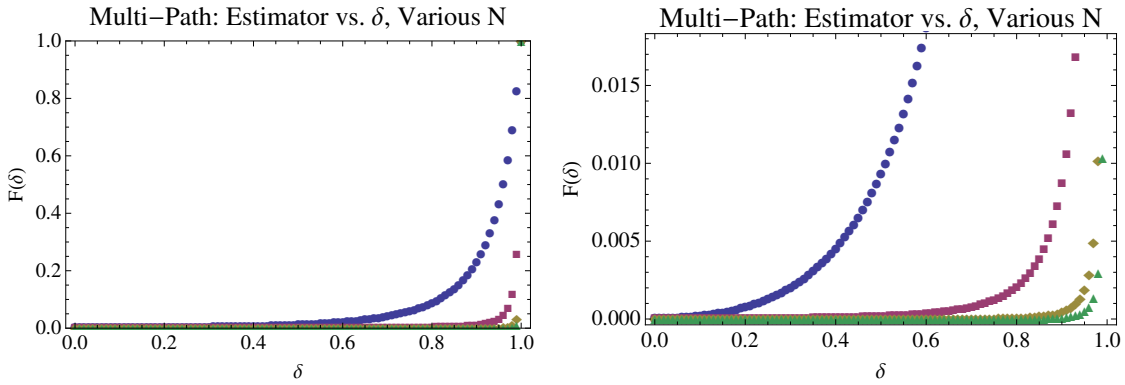



Figure 3.6: We see how $F(\delta)$ varies for various $N$ in the multi-path case. The circles, squares, diamonds, and triangles refer to $N = 10$, 100, 500, and 1000, respectively. The first plot shows the difference on the entire range of $F$ while the second is on a more relevant scale.

## 3.3 Insights from Perturbation Theory

To foster a better understanding of the possible phase transition in the multi-path undirected fragmentation, we would like to obtain the explicit dependence of $F$ on $\delta$ and $N$ in that case. If we restrict our attention to small $\delta$, we have the tools of perturbation theory at our disposal. We will cast our perturbed matrix Eq. (3.11) as another matrix with known eigenvalues and eigenstates and calculate the correction to the steady state as a function of our perturbation parameter. Then we can substitute the corrected steady state into a modified distribution estimator to get a sense for how $F$ varies with the given parameters. What follows will only apply to small fluctuations from the power law, but it will confirm the general behavior we saw in the previous section.

### 3.3.1 First-Order Perturbation Theory

As perturbation theory is often studied in connection with quantum mechanics (see [18] for a detailed treatment of time-independent perturbation theory), we adopt Dirac notation for our state vectors.

Recall that our perturbed matrix $\mathbf{M}(\delta)$ is given by Eq. (3.11). Written in this way, the unperturbed matrix is $\mathbf{M_0}$, corresponding to the power law. We claim that we know the eigenvalues $\lambda_n$ and eigenvectors $|v_n\rangle$ of $\mathbf{M_0}$. The exact eigenvalues and eigenvectors of our perturbed matrix are a power-series in $\delta$:

$$\begin{aligned} \lambda_n &= \lambda_n^{(0)} + \delta\lambda_n^{(1)} + \delta^2\lambda_n^{(2)} + \dots \\ |v_n\rangle &= |v_n^{(0)}\rangle + \delta|v_n^{(1)}\rangle + \delta^2|v_n^{(2)}\rangle + \dots \end{aligned} \tag{3.15}$$

In the following calculations, we only consider first-order corrections, which suffices for our purposes.

Then the first order correction to an eigenvalue $\lambda$ corresponding to the eigenstate $|v_0\rangle$ is given by

$$\lambda^{(1)} = \langle v_0^{(0)}| - \mathbf{M_{up}}|v_0^{(0)}\rangle, \tag{3.16}$$

where the superscript is (0) for an unperturbed value or vector and (1) for a first-order correction. The correction to the eigenvector $|v_0\rangle$ is

$$|v_0^{(1)}\rangle = \sum_{k=1}^{N-1} |v_k^{(0)}\rangle \frac{\langle v_k^{(0)}| - \mathbf{M_{up}}|v_0^{(0)}\rangle}{\lambda_0^{(0)} - \lambda_k^{(0)}}, \tag{3.17}$$

with the $|v_k^{(0)}\rangle$ and $\lambda_k^{(0)}$ normalized eigenvectors and eigenvalues of $\mathbf{M_0}$. Then the corrected eigenvector we desire is

$$|v_0\rangle = |v_0^{(0)}\rangle + \delta|v_0^{(1)}\rangle. \tag{3.18}$$

Now we must simply determine the eigenvectors and eigenvalues of our unperturbed system.

### 3.3.2 Multi-Path Fragmentation

First, we will utilize a clever change of variables for our transition matrices. We introduce $s_j = n_j x_j$. Whereas $n_j$ counted the number of pieces of a given size, $s_j$ essentially counts the amount of "stuff" represented by each piece size. By comparing $s_j$ for different $j$, we can directly compare what fraction of our conserved quantity is stored in pieces of different sizes. To transform our existing transition matrices, we simply multiply each element $(\mathbf{M})_{jk}$ by $\frac{x_j}{x_k}$. Note this leaves the diagonals unchanged.

If we denote our transformed transition matrix as $\tilde{\mathbf{M}}$, the analog to the master equation Eq. (3.2) is given by

$$\frac{ds_j}{dt} = \sum_k \tilde{\mathbf{M}}_{jk} s_k, \tag{3.19}$$

which follows both intuitively and through manipulation of Eq. (3.2). Our new conservation relation is $\sum_j s_j = X$, so in the spirit of Eq. (3.3) we have

$$\frac{d}{dt}(X) = 0 = \sum_j \frac{ds_j}{dt} = \sum_{jk} \tilde{\mathbf{M}}_{jk} s_k. \tag{3.20}$$

To make sure this holds for all $s_k$, we must have the sum of each column of $\tilde{\mathbf{M}}$ equals 0.

With the new relations established, we see that power law behavior in this scheme is given by uniform $s_j$. Now, let's examine our new multi-path undirected transition matrix, for a system of $N$ possible piece sizes:

$$\tilde{\mathbf{M}}_{\mathbf{0}} = \begin{pmatrix} -(N-1) & 1 & 1 & \dots & 1 \\ 1 & -(N-1) & 1 & \dots & 1 \\ 1 & 1 & -(N-1) & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & -(N-1) \end{pmatrix}. \tag{3.21}$$

A quick calculation shows that this matrix satisfies Eq. (3.20), so this is a valid transition matrix. From it we can determine the new $\tilde{\mathbf{M}}_{\mathbf{up}}$:

$$\tilde{\mathbf{M}}_{\mathbf{up}} = \begin{pmatrix} -(N-1) & 0 & 0 & \dots & 0 \\ 1 & -(N-2) & 0 & \dots & 0 \\ 1 & 1 & -(N-3) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 \end{pmatrix}. \tag{3.22}$$

We claim to know the eigenvectors and eigenvalues of $\tilde{\mathbf{M}}_{\mathbf{0}}$. Consider the vector $|S\rangle = \frac{1}{\sqrt{N}}(1, 1, \ldots, 1)^T$, a column vector consisting entirely of 1 entries, normalized such that the norm of the vector is 1. We can take the outer product of $|S\rangle$ with itself (note that the conjugate transpose of a real-valued vector is just the transpose). We see

$$|S\rangle\langle S| = \frac{1}{N}\begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}. \tag{3.23}$$

Now, each entry in $\tilde{\mathbf{M}}_{\mathbf{0}}$ contains a $+1$. That is, we can break up $\tilde{\mathbf{M}}_{\mathbf{0}}$ into two matrices, one made entirely of 1s and one with $-N$ along the diagonal:

$$\tilde{\mathbf{M}}_{\mathbf{0}} = \begin{pmatrix} -N & 0 & 0 & \ldots & 0 \\ 0 & -N & 0 & \ldots & 0 \\ 0 & 0 & -N & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -N \end{pmatrix} + \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & 1 & \ldots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \ldots & 1 \end{pmatrix}. \tag{3.24}$$

We recognize the first matrix as $-N$ times the $N \times N$ identity matrix and the second as $N$ times $|S\rangle\langle S|$. Thus

$$\tilde{\mathbf{M}}_{\mathbf{0}} = N\left(|S\rangle\langle S| - \mathbf{I}\right). \tag{3.25}$$

This observation is key, as we know the eigenvectors for $|S\rangle\langle S|$. Those are the eigenvectors of $\tilde{\mathbf{M}}_{\mathbf{0}}$ as well, but with the eigenvalues shifted by $-N$ due to the identity matrix. The eigenvectors are $|S\rangle$ (eigenvalue 0) and $N-1$ eigenvectors orthogonal to $|S\rangle$ (all eigenvalue $-N$). An eigenvalue of 0 means $|S\rangle$ is our steady-state.

### 3.3.3 Perturbation Corrections

Our first task is to confirm that our steady state stays steady. That is, we want the first order correction to our eigenvalue to be zero. Let $\lambda_S$ denote the eigenvalue of $|S\rangle$. From Eq. (3.16),

$$\begin{aligned} \lambda_S^{(1)} &= \langle S^{(0)}| - \tilde{\mathbf{M}}_{\mathbf{up}}|S^{(0)}\rangle \\ &= \frac{1}{N}(1, 1, \ldots, 1)\begin{pmatrix} N-1 & 0 & \ldots & 0 \\ 1 & N-2 & \ldots & 0 \\ 1 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \ldots & 0 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N}\left(N + (N-2) + \cdots - (N-2) - N\right) \\
&= 0. \qquad (3.26)
\end{aligned}
$$

Hence the eigenvalue correction is zero, so this eigenvalue does not change. Our steady state stays steady.

Now we compute the first-order correction to $|S\rangle$. From Eq. (3.17), we have

$$
|S^{(1)}\rangle = -\frac{1}{N}\sum_{k=1}^{N-1} |v_k^{(0)}\rangle\langle v_k^{(0)}|\tilde{\mathbf{M}}_{\mathbf{up}}|S\rangle, \qquad (3.27)
$$

as $\lambda_S - \lambda_k = -N$ for all $k$, so we can factor it out. We also factored out a negative sign. We can easily calculate $\tilde{\mathbf{M}}_{\mathbf{up}}|S\rangle$. Recall that for a given matrix with eigenstates, we can decompose an arbitrary vector $|v\rangle$ into a sum of inner products and eigenstates [18]:

$$
|v\rangle = \sum_j |v_j\rangle\langle v_j|v\rangle, \qquad (3.28)
$$

where $j$ sums over all the eigenstates of our matrix. We can use this fact to greatly simplify our correction. Define

$$
|v\rangle = \tilde{\mathbf{M}}_{\mathbf{up}}|S\rangle. \qquad (3.29)
$$

We decompose $|v\rangle$ into a sum involving the eigenstates of $\tilde{\mathbf{M}}_{\mathbf{0}}$. The last piece of the puzzle is to recognize from above that because $\lambda_S^{(0)} = 0$, $|S^{(0)}\rangle\langle S^{(0)}|v\rangle = 0$. Thus we may add it to our existing expressions without changing their value. Then

$$
\begin{aligned}
|S^{(1)}\rangle &= -\frac{1}{N}\sum_{k=1}^{N-1} |v_k^{(0)}\rangle\langle v_k^{(0)}|v\rangle \\
&= -\frac{1}{N}\sum_{k=1}^{N-1} |v_k^{(0)}\rangle\langle v_k^{(0)}|v\rangle + |S^{(0)}\rangle\langle S^{(0)}|v\rangle \\
&= -\frac{1}{N}\sum_{k=0}^{N-1} |v_k^{(0)}\rangle\langle v_k^{(0)}|v\rangle \\
&= -\frac{1}{N}|v\rangle \\
&= \frac{1}{\sqrt{N}}\left(1 - \frac{1}{N}, 1 - \frac{3}{N}, \ldots, -\left(1 - \frac{3}{N}\right), -\left(1 - \frac{1}{N}\right)\right)^T. \qquad (3.30)
\end{aligned}
$$

Note that in the above we represented $|S\rangle$ as $|v_0\rangle$ in the sum. In the final correction vector, the terms proportional to $1/N$ are such that the vector elements

sweep from $+1$ to $-1$ in steps of $2/N$. If $N$ is odd, then we expect 0 to appear in the vector. Put another way, the $j$th element of $|S^{(1)}\rangle$ is given by $1 - \frac{2j-1}{N}$, multiplied by a normalizing factor. Hence the $j$th element of the perturbed steady state vector is given by, up to normalization,

$$|S\rangle_j = 1 + \delta\left(1 - \frac{2j-1}{N}\right). \tag{3.31}$$

### 3.3.4 Modified Distribution Estimator

Now that we have our steady state, we wish to modify the distribution estimator, Eq. (3.14), to work in our new conservation scheme. The power law distribution is now given by $X/N$, and we normalize $|S\rangle$ so that $\sum s_j = 1$ (this parallels the construction of $F$ in Section 3.2.1). Then given an arbitrary steady state, $F$ takes the form

$$F(\delta) = \frac{\sum_{j=1}^{N}\left|s_j - \frac{1}{N}\right|^2}{\left|1 - \frac{1}{N}\right|^2 + \sum_{j=2}^{N}\left|\frac{1}{N}\right|^2} = \frac{N\sum_{j=1}^{N}\left|s_j - \frac{1}{N}\right|^2}{N-1}. \tag{3.32}$$

We must calculate the sum in the numerator. As we have a formula for each element of $|S\rangle$, we can do this easily. Note that we normalize our steady state by an overall factor of $1/N$ to satisfy the conservation relation with $X = 1$. Then

$$\begin{aligned}
\sum_{j=1}^{N}\left|s_j - \frac{1}{N}\right|^2 &= \sum_{j=1}^{N}\left|\frac{1}{N}\left(1 + \delta\left(1 - \frac{2j-1}{N}\right)\right) - \frac{1}{N}\right|^2 \\
&= \sum_{j=1}^{N}\left|\frac{1}{N}\delta\left(1 - \frac{2j-1}{N}\right)\right|^2 \\
&= \frac{\delta^2}{N^4}\sum_{j=1}^{N}(N - 2j + 1)^2 \\
&= \frac{\delta^2}{N^4}\left(\frac{1}{3}N(N^2-1)\right) \\
&= \frac{\delta^2}{3N^3}(N^2-1).
\end{aligned} \tag{3.33}$$

We substitute this into Eq. (3.32) to obtain, after some manipulation,

$$F(\delta) = \delta^2\frac{N+1}{3N^2}. \tag{3.34}$$

So as $N$ grows and $\delta$ is fixed, our estimator varies as $1/N$. Hence, as our matrix size increases, we expect to see $F$ approach 0 for small $\delta$. This is exactly the

behavior we witnessed in the numerical simulations. Moreover, this holds for any set of possible piece sizes, as we did not assume any particular set, so it is a general feature of the process itself.

# Chapter 4

# Conclusion

In this work, we have considered how a conserved quantity, under fragmentation, leads to power law and Benford behavior. We have found that not only do a wide variety of fragmentation processes result in a power law distribution of piece sizes, but knowing the magnitude of the initial conserved quantity $X$ and the number of possible piece sizes $N$, we can reasonably predict whether a list of fragment sizes will be Benford. Here we reflect on our findings and consider possible avenues for future research.

Our initial goal was to qualify Lemons' result. Our integer fragmentation model constrained $X$ to be much larger then $N^2 x_N$. Meanwhile, our canonical ensemble model collapsed to the Lemons result for $X \gg N x_N$. While we have numerical and some analytical evidence, for these limits, our rigorous bounds for $\langle n_j \rangle$ do a poor job in comparison to the approximation or canonical ensemble methods. An important task, for which we are confident, is to tighten those bounds. We note that $P_H(X)$ shows up in other mathematical investigations. It would also be of great use to derive a more accurate closed-form expression to represent it, whether that be through the bounds or an improved approximation.

Just as we used mathematical methods to place constraints on $X$, we could perform a similar analysis to place constraints on $N$. One future project is to treat the problem of Section 2.3 analytically. This result would be more exact than the rough order of magnitude threshold our numerical calculations provided. A more complicated next step would be to formulate a constraint on $N$ given an arbitrary set $H$.

In Section 3.1, we considered a special class of fragmentation processes, the single-path and multi-path directed and undirected fragmentations. We found that undirected graphs can be weighted in such a way to elicit a power law steady state. We would like to consider other potential classes of break-up protocols and perform a similar analysis. Of particular interest would be shattering. We suspect that a shattering results in a Poisson rather than power law distribution [13].

Ideally, we can conceive a generalized fragmentation model that can be adjusted to give both outcomes. With that flexibility, we could probe the behavior between the extremes.

In the final section, Section 3.3, we showed that perturbation theory confirms the numerical results of Section 3.2 for the multi-path undirected fragmentation. We would like to demonstrate for the single-path case as well. Perhaps the most interesting result of our investigations was the existence of a phase transition in the multi-path undirected fragmentation steady state and the robustness of the power law that results. We would like to explore this property further, perhaps looking at parallels in real processes.

While our work is theoretical in nature, we hope that what we discover will be of use in dealings with the real world. Our results may lead to reliable methods for determining when we should expect data sets to exhibit Benford behavior. Finally, when all is said and done, we hope that the investigations stemming from this work and that affiliated are not a small thing in the world but a very large thing indeed.

# Appendix

## A Section 2.2.2 - Derivation of Equation (2.27)

We wish to evaluate Eq. (2.12),

$$\langle n_j \rangle = \frac{1}{P_H(X)} \sum_{i=0}^{\lfloor X/x_j \rfloor} P_H(X - ix_j),$$

for $\langle n_j \rangle$. We use the approximation for $P_H(X)$ given by Eq. (2.26):

$$P_H(X) \approx \frac{X^{N-1}}{(N-1)!D_N} + \frac{X^{N-2}}{2(N-2)!D_N}\left(x_2 + \sum_{l=2}^{N} x_l\right).$$

For ease of calculation, assume $x_j | X$ and let $\gamma = \left(x_2 + \sum_{l=2}^{N} x_l\right)$. Then we see that

$$\begin{aligned}
\frac{1}{P_H(X)} &= \frac{1}{\frac{X^{N-1}}{(N-1)!D_N} + \frac{\gamma X^{N-2}}{2(N-2)!D_N}} \\
&= \frac{1}{\left(\frac{2X^{N-1} + \gamma(N-1)X^{N-2}}{2(N-1)!D_N}\right)} \\
&= \frac{2(N-1)!D_N}{2X^{N-1} + \gamma(N-1)X^{N-2}} \\
&= \frac{(N-1)!D_N}{X^{N-1}\left(1 + \frac{\gamma(N-1)}{2X}\right)}.
\end{aligned}$$

We use this result in Eq. (2.12):

$$
\begin{aligned}
\langle n_j \rangle &= \frac{(N-1)!D_N}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \sum_{i=1}^{X/x_j} \left[ \frac{(X-ix_j)^{N-1}}{(N-1)!D_N} + \frac{\gamma(X-ix_j)^{N-2}}{2(N-2)!D_N} \right] \\
&= \frac{(N-1)!D_N}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \sum_{k=0}^{X/x_j-1} \left[ \frac{(kx_j)^{N-1}}{(N-1)!D_N} + \frac{\gamma(kx_j)^{N-2}}{2(N-2)!D_N} \right] \\
&= \frac{(N-1)!D_N}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \left[ \frac{x_j^{N-1}}{(N-1)!D_N} \sum_{k=0}^{X/x_j-1} k^{N-1} + \frac{\gamma x_j^{N-2}}{2(N-2)!D_N} \sum_{k=0}^{X/x_j-1} k^{N-2} \right] \\
&= \frac{1}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \left[ x_j^{N-1} \sum_{k=0}^{X/x_j-1} k^{N-1} + \frac{1}{2}\gamma\,(N-1)\,x_j^{N-2} \sum_{k=0}^{X/x_j-1} k^{N-2} \right] \\
&\approx \frac{1}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \left[ \frac{x_j^{N-1}}{N} \left(\frac{X}{x_j}-1\right)^N + \frac{1}{2}\gamma\,(N-1)\,\frac{x_j^{N-2}}{N-1} \left(\frac{X}{x_j}-1\right)^{N-1} \right] \\
&= \frac{1}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \left[ \frac{x_j^{N-1}}{N} \left(\frac{X}{x_j}\right)^N \left(1-\frac{x_j}{X}\right)^N + \right. \\
&\quad \left. \frac{1}{2}\gamma x_j^{N-2} \left(\frac{X}{x_j}\right)^{N-1} \left(1-\frac{x_j}{X}\right)^{N-1} \right] \\
&= \frac{1}{X^{N-1}\left(1+\gamma\frac{(N-1)}{2X}\right)} \frac{X^N}{Nx_j} \left[ \left(1-\frac{x_j}{X}\right)^N + \frac{\gamma N}{2X}\left(1-\frac{x_j}{X}\right)^{N-1} \right] \\
&= \frac{X}{Nx_j}\left(1+\gamma\frac{(N-1)}{2X}\right)^{-1} \left[ \left(1-\frac{x_j}{X}\right)^N + \frac{\gamma N}{2X}\left(1-\frac{x_j}{X}\right)^{N-1} \right] \\
&= \frac{X}{Nx_j}\left(1+\gamma\frac{(N-1)}{2X}\right)^{-1} \left(1-\frac{x_j}{X}\right)^N \left[ 1 + \frac{\gamma N}{2X}\left(1-\frac{x_j}{X}\right)^{-1} \right]. \qquad (1)
\end{aligned}
$$

# B Section 2.2.3 - Evaluation of Equation (2.30)

We wish to evaluate the lower bound for $P_H(X)$, Eq. (2.30):

$$
P_H(X) > \int_{n_N=0}^{L_N} \int_{n_{N-1}=0}^{L_{N-1}} \ldots \int_{n_2=0}^{L_2} dn_2 \cdots dn_N, \qquad (2)
$$

where the $L_j$ are defined by Eq. (2.18). We use induction. We see that

$$\begin{aligned} P_H(X) &> \int_{n_N=0}^{L_N} \int_{n_{N-1}=0}^{L_{N-1}} ... \int_{n_2=0}^{L_2} dn_2 \cdots dn_N \\ &= \int_{n_N=0}^{L_N} \int_{n_{N-1}=0}^{L_{N-1}} ... \int_{n_3=0}^{L_3} L_2 \; dn_3 \cdots dn_N \\ &= \int_{n_N=0}^{L_N} \int_{n_{N-1}=0}^{L_{N-1}} ... \int_{n_3=0}^{L_3} \frac{x_3}{x_2}(L_3 - n_3) \; dn_3 \cdots dn_N. \end{aligned} \tag{3}$$

We claim that after integrating over $n_k$, for all $k \leq N$, Eq. (2) becomes

$$P_H(X) > \int_{n_N=0}^{L_N} ... \int_{n_{k+1}=0}^{L_{k+1}} \frac{x_k^{k-2}}{(k-1)!D_{k-1}} L_k^{k-1} \; dn_{k+1} \cdots dn_N, \tag{4}$$

with $D_k$ given by Eq. (2.21),

$$D_k = \prod_{j=1}^{k} x_j.$$

Our base case is $k = 2$, Eq. (3). We show that our hypothesis holds for $k+1$. Recall the recursion relation on $L_k$ given by Eq. (2.18):

$$L_k = \frac{x_{k+1}}{x_k}(L_{k+1} - n_{k+1}), \qquad k = \{2, 3, \ldots, N-1\}.$$

We perform an integration on Eq. (4):

$$\begin{aligned} & \int_{n_N=0}^{L_N} ... \int_{n_{k+1}=0}^{L_{k+1}} \frac{x_k^{k-2}}{(k-1)!D_{k-1}} L_k^{k-1} \; dn_{k+1} \cdots dn_N \\ = & \int_{n_N=0}^{L_N} ... \int_{n_{k+1}=0}^{L_{k+1}} \frac{x_k^{k-2}}{(k-1)!D_{k-1}} \left(\frac{x_{k+1}}{x_k}\right)^{k-1} (L_{k+1} - n_{k+1})^{k-1} \; dn_{k+1} \cdots dn_N \\ = & \int_{n_N=0}^{L_N} ... \int_{n_{k+1}=0}^{L_{k+1}} \frac{x_{k+1}^{k-1}}{(k-1)!D_k} (L_{k+1} - n_{k+1})^{k-1} \; dn_{k+1} \cdots dn_N \end{aligned}$$

$$
\begin{aligned}
&= \int_{n_N=0}^{L_N} \cdots \int_{n_{k+2}=0}^{L_{k+2}} \frac{x_{k+1}^{k-1}}{(k-1)!D_k} \frac{1}{k} L_{k+1}^k \; dn_{k+2} \cdots dn_N \\
&= \int_{n_N=0}^{L_N} \cdots \int_{n_{k+2}=0}^{L_{k+2}} \frac{x_{k+1}^{k-1}}{k!D_k} L_{k+1}^k \; dn_{k+2} \cdots dn_N, \qquad (5)
\end{aligned}
$$

which is the induction result for $k+1$. Thus our induction holds, and after we integrate through $n_N$, we are left with

$$
\begin{aligned}
\frac{x_N^{N-2}}{(N-1)!D_{N-1}} L_N^{N-1} &= \frac{x_N^{N-2}}{(N-1)!D_{N-1}} \left( \frac{X}{x_N} \right)^{N-1} \\
&= \frac{X^{N-1}}{(N-1)!D_N}. \qquad (6)
\end{aligned}
$$

This is Eq. (2.31), as desired.

# References


[1] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 9th edition, 1964.

[2] George E. Andrews. *The Theory of Partitions*. Addison-Wesley Publishing Company, 1976.

[3] Matthias Beck, Ira M. Gessel, and Takao Komatsu. The polynomial part of a restricted partition function related to the frobenius problem. *The Electronic Journal of Combinatorics*, 8(1), 2001.

[4] Thealexa Becker, Taylor C. Corcoran, Alec Greaves-Tunnell, Joseph R. Iafrate, Joy Jing, Steven J. Miller, Jaclyn D. Porfilio, Ryan Ronan, Jirapat Samranvedhya, and Frederick W. Strauch. Benford's law and continuous dependent random variables. Preprint, 2013.

[5] Frank Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, March 1938.

[6] Arno Berger and Theodore P. Hill. Benford's law strikes back: No simple explanation in sight for mathematical gem. *The Mathematical Intelligencer*, 33(1):85–91, 2011.

[7] K. C. Chase and A. Z. Mekjian. Nuclear fragmentation and its parallels. *Physical Review C*, 49:2164, 1994.

[8] Thomas H. Corman, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, 3rd edition, 2009.

[9] Theodore P. Hill. Base-invariance implies benford's law. *Proceedings of the American Philosophical Society*, 123(3):887–895, March 1995.

[10] Theodore P. Hill. A statistical derivation of the significant-digit law. *Statistical Science*, 10(4):354–363, November 1995.

[11] Don S. Lemons. On the numbers of things and the distributions of first digits. *American Journal of Physics*, 54:816, 1986.

[12] Jason Long. Testing benford's law. `http://testingbenfordslaw.com/`. Accessed May 4, 2014.

[13] Steven J. Miller and Ramin Takloo-Bighash. *An Invitation to Modern Number Theory*. Princeton University Press, 2006.

[14] Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of Mathematics*, 4(1):39–40, 1881.

[15] Mark Newman. *Networks: An Introduction*. Oxford University Press Inc., New York, 2010.

[16] R. S. Pinkham. On the distribution of first significant digits. *The Annals of Mathematical Statistics*, 32:1223, 1961.

[17] Daniel V. Schroeder. *An Introduction to Thermal Physics*. Addison-Wesley, 1999.

[18] John S. Townsend. *A Modern Approach to Quantum Mechanics*. University Science Books, 2nd edition, 2012.

[19] Nicolaas Godfried van Kampen. *Stochastic Processes in Physics and Chemistry*. North Holland, 3rd edition, 2007.

[20] Charles G. Wohl. Benford's law. Unpublished.